The London School of Economics and Political Science

Predictability and the decay of information in mathematical and physical systems

Ewelina Sienkiewicz

A thesis submitted to the Department of Statistics
of the London School of Economics for the degree of Doctor of Philosophy.

London, December, 2017

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of less than 100,000 words.

Statement of conjoint work

I confirm that chapter 7 was jointly co-authored with Dr Erica L Thompson and I contributed to 95% of this work.

Abstract

This thesis explores the predictability of nonlinear systems, both mathematical systems (as realised on a digital computer) and geophysical systems (the El Niño Phenomena, and climate prediction via downscaling). How far into the future does a forecast system provide information beyond that available purely from the past? How does information in a probabilistic forecast decay with time? Is it true that, no matter how good the simulation model used for prediction is, there will be a point where predictability is lost? That is, that there is always a time horizon beyond which any forecast fails to yield useful information.

The two main limits to predictability are identified and discussed. Sensitivity to initial condition complicates the forecasting of chaotic dynamical systems, even when the model is perfect. Structural model error (model inadequacy) is a distinct cause of the decay of predictability, a decay that may often be mistakenly interpreted as resulting from chaos. These features are distinguished and demonstrated both in low-dimensional mathematical systems and weather and climate models.

Model inadequacy is shown to be important in real-world forecasting, with reference to Columbia University's C-Z model for El Niño predictions and climate models used in the North American Regional Climate Change Assessment Program (NARCCAP). Repercussions for forecast performance are discussed. In short, (i) NARCCAP regional simulations are quickly inconsistent with the global simulations used to drive them, (ii) the C-Z model allows experiments into the decay of predictability when one model version is employed as the system, and a second, structurally distinct model version is used as the model.

The decay of predictability is studied from the view point of information theory. Information theoretic tools are allied both to mathematical system-model pairs and to physical system-model pairs. A flaw in formulating one such tool, proposed by Du and Smith (2012, PRE) is exposed and alternative normalisations are explored in various experiments. A quantity called the information deficit, introduced in that same paper, is considered in several settings. New properties of the information deficit are discovered, and it is demonstrated that the information deficit can be a useful tool in identifying (and correcting) shortcomings of a forecasting system.

Acknowledgements

I would like to express my gratitude to all the wonderful people mentioned here. They supported me in my research over the last few years and helped me bring the PhD thesis to the finish line.

To my supervisor Professor Leonard Smith, thank you for giving me directions and guidance and for sharing your unlimited enthusiasm and wisdom. You made my PhD studies exciting and inspiring.

To Dr Erica Thompson and Dr Edward Wheatcroft, it has been a pleasure working with you. Thank you for all your greatly appreciated help and advice and for being so generous with your time.

To my husband Konrad for being supportive, positive and fulfilling a huge list of day-to-day duties exceedingly well.

Finally, to my mother and little daughter Nell for bringing lots of love and joy to every day.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The main theme of this thesis is predictability and the exploration of useful information in nonlinear forecasting situations. Two distinct factors that limit predictability of nonlinear models, specifically structural model error and uncertainty in the initial condition, are explored in various experiments. We consider predictability from the perspective of information theory. Throughout the thesis, models are used to form probabilistic forecast distributions and information is quantified in terms of information (measured in bits) with reference to some baseline. Climatology, which comprises of typical values derived from a large number of past observations of the system, serves as this reference standard.

This chapter provides an overview of the thesis. Some terminologies may be new to the reader and are explained in later chapters when they are used for the first time. In short, the thesis is structured as follows: chapter 2 and chapter 3 provide background knowledge used and explored throughout the thesis. Broadly, the mathematical content can be found in chapter 4 and 5, while applications to two

real-world modelling systems are described in chapter 6 (the Cane-Zebiak [21] model of El Niño) and chapter 7 (the set of NARCCAP models of the Earths climate, focused on North America [6]).

Chapter 2 presents the foundations of nonlinear dynamics that we will adopt. Methodology underpinning prediction scenarios used later in the thesis such as ensemble forecasting, are introduced. There are no novel contributions in chapter 2 beyond the presentation of background material.

Chapter 3 introduces a number of measures for quantifying predictability which are exploited and developed throughout this thesis. We discuss the concept of predictability and how it decays, define the procedures required to convert ensembles into probability density functions and describe the methods of model selection used in the thesis. While there is no new theory introduced in this chapter, a distinction between two methods of computing the Ignorance and the Information Deficit in sections 3.3 and 3.4 is new to this thesis. Several inconsistencies found in the literature are identified and an improved approach to further calculations is proposed.

Chapter 4 clarifies the distinction between model inadequacy and sensitivity to initial condition uncertainty. Structural model error occurs when the equations of the model differ from those of the system that generated the data. In the real-world, this is always the case. A system is sensitive to initial conditions when small differences cause a large difference in how the system evolves into the future (under the same dynamics). Most notably, initial condition error is caused by observational uncertainty which is always likely to be present in real-world forecasting scenarios. This means that exact predictions of the future would be impossible even if the true nature of the system dynamics were known perfectly [95]. New results in this chapter show that, if the mathematical structure of the equations of the system

differ from those of the model, predictability decays in a different way to when uncertainty is caused purely by initial condition error. Another novel result of this chapter lies in illustrating how the predictability of a chaotic system varies with the date on which the forecast is launched (that is the location of the ensemble of initial conditions). Probabilistic prediction can be improved by taking more precise observations (decreasing the noise level), more expensive forecasts (increasing the number of Monte-Carlo ensembles) or using some data assimilation technique [100, 98, 101]. None of these actions can improve errors related to model inadequacy [103]. More generally, acknowledging and recognising the distinction between model inadequacy and sensitivity to initial condition can lead to more effective resource allocation in dealing with these two limits of predictability.

Chapter 5 expands our understanding of the decay of predictability focusing on low-dimensional mathematical system-model pairs. Experiments in this chapter are focused on the Logistic and Quartic maps as an illustrative system-model pair. Whilst the Quartic Map was introduced previously [38], its properties are more fully analysed for the first time in this thesis. Using Relative Entropy, we derive a threshold level of imperfection in the model at which the model and system are distinguishable. Bifurcation diagrams of the model and the system support this conclusion. The decay of predictability is contrasted in both a perfect and an imperfect model scenario. The Information Deficit, a tool to identify and improve predictability, is used here for the first time. In this context, it is shown that one of the two normalisation methods in the literature (introduced in chapter 3) is superior to the other, and the reasons why are presented. The novel contributions in this chapter form a significant contribution to a paper in preparation for submission [101].

Chapter 6 focuses on study of the Cane-Zebiak model [21], a real-world operational

model developed at Columbia University in New York. This is an example of a very complex model used to forecast events of social and economic interest such as El Niño [42]. El Niño is a global climatic phenomenon with widespread socio-economic impacts. Prediction of El Niño behaviour would be of great value to society and would have the potential to reduce loss and damage. Existing forecast methods are inadequate to provide useful information in the timescales of interest. This provided the motivation to study predictability in the context of the C-Z model. Our original research goals were: (i) to introduce a structural difference into the computer code of the model, and then define this alternative code as the 'system' for which the C-Z model would serve as a structurally imperfect model. We would then proceed in a similar way as with the lower-dimensional system-model pairs. (ii) to construct several structurally distinct versions of the C-Z model and examine the impact of using these in forecasting when the total computational resource was fixed.

In this chapter, we analyse both the qualitative and quantitative behaviour of the C-Z model, demonstrate that it incorporates the features of El Niño realistically and show it exhibits sensitivity to small differences in the initial conditions. While trying to obtain structurally different versions of the C-Z model, it soon became apparent that the numerical model itself is fragile. Small changes in model structure would result in very different qualitative behaviour, as discussed briefly in chapter 6. As a result, the research focused on contrasting a simple structural difference: the system being the C-Z model with the standard integration time step of 10 days, while the model is the same computer code with the time step set to 5 days. The decay of predictability in the Perfect Model Scenario was studied in detail for the first time. Its decay in this system-model pair was then considered. To our knowledge, this is the first time an imperfect model scenario of this nature has been performed on

an operational El Niño forecasting model. The Information Deficit again helped to reveal practical difficulty in the experimental methodology, such as how to define outcomes and ensembles from a set of model runs, specifically that an outcome should be chosen randomly.

Chapter 7 reports the initial steps of another experiment where the quality of the results we unearthed led to the early rejection of further research plans. The NARCCAP [6] data set is a large database in which regional climate models with high spatial resolution outputs are simulated by lower resolution models that are global in scope. In NARCCAP, regional models are usually driven by a global model in a one-way fashion; no information from the regional model feeds back into the evolution of the global model that drives it. The result of this could be that the physical fields (temperature, rainfall, etc) simulated by the two types of models will diverge with time, leading to physical inconsistency between them. At that point, the use of these simulations in science-informed policy making would be questionable. Decision makers exploit available outputs from regional models as an aid to policy development. If they deviate too much from the global model that simulated them this could lead to poor decision making. The initial experimental design was to explore this divergence over time and develop a test of internal consistency in order to determine whether computer model simulations are reliable or not. Results shown in this chapter revealed that significant divergence was almost immediate: the regional model 'climate' is inconsistent with the global model 'climate' in a fundamental sense. This made studying their divergence over time impossible but lead to additional work with Dr Erica Thompson and a paper is in preparation on this fundamental challenge to the NARCCAP approach [108].

A summary, conclusions and the identification of new material in each chapter are

included at the end of that chapter.

Chapter 2

Introduction to Predictability of Dynamical Systems

This chapter presents a literature review of relevant concepts in the theory of dynamical systems, which are a subject of the exploration in the thesis. There is no new material in this chapter beyond the presentation.

This chapter is structured as follows. First, in sections 2.1 - 2.3 terminology and properties of nonlinear dynamical systems are defined. In section 2.4 the concept of model and system is described. Section 2.5 defines two scenarios used in experiments later in the thesis: the Perfect and Imperfect Model Scenarios. Section 2.6 gives a brief introduction to ensemble forecasting and the techniques applied in the formation of an ensemble.

2.1 Nonlinear dynamical systems

Most processes in nature are best modelled as **nonlinear**, i.e. change of the output is not proportional to change of the input [16]. In this thesis we are going to focus on nonlinear systems, which are typically described by nonlinear differential equations or nonlinear discrete-time mappings.

A **dynamical system** [110], is a mathematical system whose evolution in time from some initial state is governed by set of rules. The evolution of the dynamical system is best described in a **state space** $\mathbb{S}$[1]. A dynamical system can be defined in the form $x_t = F^t(x_0)$, where $x \in \mathbb{S}$. F is a function (including any parameters) which defines the dynamics of the system, x_0 is the **initial state** or **initial condition** of the system, t is the evolution time.

As a given initial condition of a dissipative dynamical system is iterated forward in time, it will evolve towards an **attractor** [104], a subset of points in the D-dimensional state space which are explored by the system as $t \to \infty$. One example of an attractor, the Ikeda attractor, is shown in Fig.2.1. States that are inconsistent with the long-term behaviour of the dynamical system are referred to as **transient** ($TRAN$). Typically, the trajectory of a randomly chosen initial conditions at an early stage of the evolution, is far from the attractor. In this thesis, we assume that a system has been evolving long enough that there are effectively no transient states. The same can not be said for the model. The **trajectory** [107] consists of the ordered set of future positions in state space of a point that is evolved in time.

[1]For example $\mathbb{S} \equiv \mathbb{R}^D$.

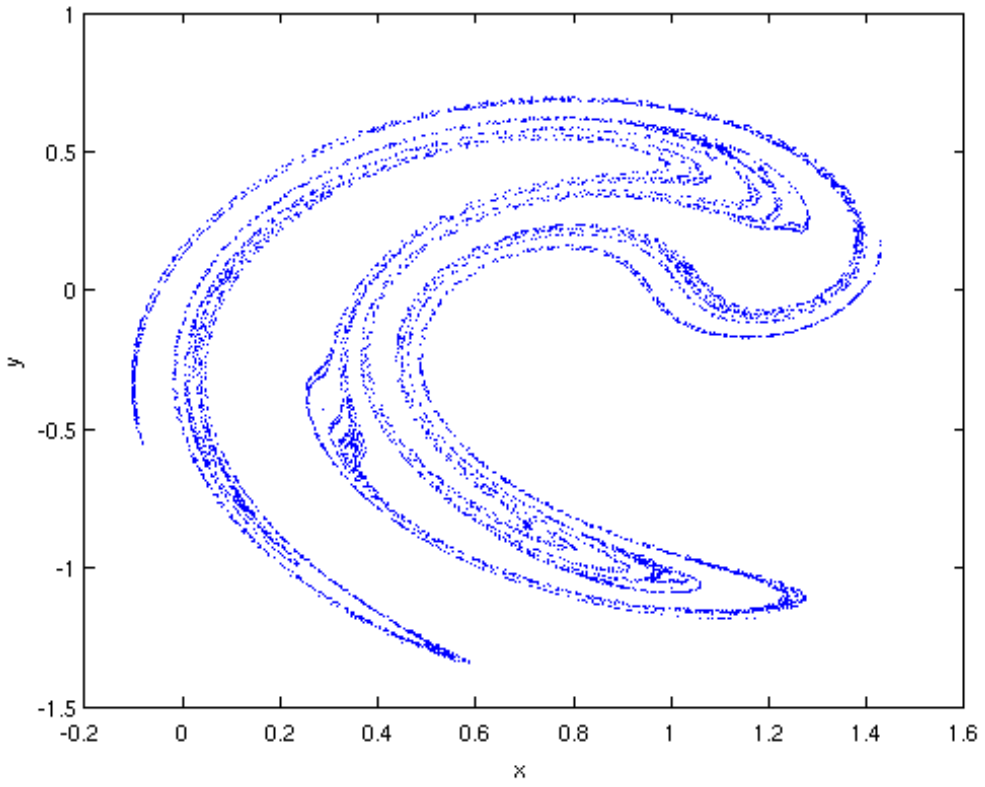


Figure 2.1: The Ikeda attractor as defined in Appendix A.3. In the long term, all initial conditions will converge to this attractor.

Dynamical systems in continuous time described by differential equations for dx/dt in terms of x are referred to as **flows** [105]. Often, there is no analytical solution to the ordinary nonlinear differential equations; nevertheless, the continuous dynamical system can be simulated using some numerical integration scheme, e.g. 4th order Runge-Kutta approximation [83]. Dynamical systems that are evolved in discrete time described by a set of differential equations for x_{n+1} in terms of x_n are called **maps**.

Two different categories of dynamical systems can be distinguished: stochastic and deterministic [18]. In a **stochastic dynamical system** some elements of the sys-

tem are random; for example, a random noise may be introduced at each time step or there may be a random variation in some parameters. By contrast, in a **deterministic dynamical system** the initial state determines all future states under iteration and a randomness does not affect evolution. In this thesis we study deterministic dynamical systems; the climate models (chapter 7) and El Niño model (chapter 6) are formulated as deterministic dynamical systems.

2.2 Climatology

Climatology has two common meanings. The first is a study of climate of an area [85], the second is to represent a distribution based on observations over a defined period of time [19], generalising this first notion to observations from any dynamical system. A climatology, describes the long-term behaviour of the dynamical system and is essentially a description of the attractor as a probability distribution in state space. In practice it is formed from past observations of a system, but excluding transient states. Climatological distributions are usually constructed using parametric or non-parametric density estimation methods.

Climatology can be used as a forecasting distribution [5]. For example, suppose that we would like to know the temperature in London in 3 months from today. If weather models do not provide much information for such a long period in advance, then one could look at historical statistics over the last 30 years for the date of interest to have an indication of what to expect. This climatology contains information about the seasonal variation of the weather. At a minimum, we hope for an operational dynamical weather forecast to perform as well as a climatological forecast, which can act as a benchmark or reference for assessing the quality of predictions.

2.3 Chaos

Some dynamical systems are chaotic. Smith [100] defines chaos as 'a mathematical dynamical system which is deterministic, recurrent and has sensitive dependence on initial state'. The sensitivity to the initial condition means that starting iterations in a slightly different position on the dynamical system, can cause drastically different behaviour in later states [65]. Another feature of a chaotic system is that it is both recurrent, that is, the system state returns infinitely often and arbitrarily closely to its original condition, and deterministic, as no randomness is involved in the evolution of the system.

2.4 System-Model pair

Making predictions of a dynamical systems requires a forecasting model. Different disciplines may use a different mathematical description to construct a model of a dynamical system. These could be differential equations in physics, logic and graphical models in artificial intelligence, sequential decision processes in operations research or a stochastic model in statistics [92] to give just a few examples.

It is not possible to describe a system accurately (that is, with a mathematical precision), as a model can only ever be an approximate representation of nature's laws. A model dynamics F tries to mathematically formalise a system's dynamics $\tilde{F}$ [93]. Model and system are subtractable, therefore they share the same state space [99]. An observation of a forecasted system state is called an **outcome** or a target. A set of model forecasts and the corresponding outcomes will constitute a **forecast-outcome archive**.

In this thesis in chapters 4 and 5, different mathematical theoretical models will

constitute both system and model, allowing theoretical concepts to be developed and demonstrated. In chapter 6, a system-model pair will concern the much more complex oceanic-atmospheric model of partial differential equations, while in chapter 7 climate models in the North American Regional Climate Change Assessment Program (NARCCAP) [6] will be used.

2.5 Perfect and Imperfect Model Scenarios

To investigate the properties of the forecasting model two cases are considered in this thesis: the Perfect Model Scenario (PMS) [95] and the Imperfect Model Scenario (IMS) [96]. Both concepts are defined below.

2.5.1 Perfect Model Scenario

In the Perfect Model Scenario (PMS) both the model and the system are mathematically equivalent $F=\tilde{F}$, that is, the dynamics of the system is described perfectly by the model. Arguably, such a situation is possible only in pure mathematics, never in reality. Under the PMS, a forecast trajectory is imperfect only when the exact initial conditions or parameter values are unknown. If there is some uncertainty in the initial condition, then there will be a corresponding uncertainty in the forecast, but the forecast range can contain the correct answer. If a probabilistic method is used, a reliable probability forecast for the outcome can in principle be constructed.

2.5.2 Imperfect Model Scenario

In the Imperfect Model Scenario (IMS) a model only approximates the dynamics of the system $F \neq \tilde{F}$. There is structural model error, that is, the mathematical form of the model differs from that of the system (if such a thing even exists). This idea

reflects real world forecasting. Processes in nature or real life can not be described precisely in the form of mathematical equations, because they are too complex. Like the Perfect Model Scenario, in the IMS the initial condition uncertainty is due to observational error [98]. In the IMS, however, there is an additional 'model uncertainty' caused by the error in the dynamics of the model itself. For short time periods this uncertainty may be negligible but over some time interval the forecast will almost surely depart from the true outcome, even in cases where all initial condition uncertainty is accounted for.

2.6 Ensemble forecasting

Ensemble forecasting [64, 109, 27] is based on running simulations of the same model with slightly different initial conditions in order to reveal the diversity of the model trajectories in hopes of quantifying the impact of the initial condition uncertainty. This approach is a Monte-Carlo method and is often used to make a probabilistic forecast of a dynamical system [63, 109, 64]. In the Perfect Model Scenario, these probabilistic forecasts will be reliable. Ensembles are widely deployed [3, 29, 71] in Numerical Weather Prediction (NWP) to give a range of possible future states of the system.

An ensemble is formed by sampling m initial conditions consistent with an observation of the initial condition of a dynamical system and, ideally, the long term dynamics of the system [97]. This constitutes an **initial condition ensemble**, while an **ensemble member** is a single initial condition of such an ensemble. Each ensemble member is then evolved forward into the future by a period of time called a **lead time**, under the forecast model, to produce an ensemble of point forecasts. In this thesis we will refer to each ensemble and corresponding outcome as an **ensemble-**

outcome pair and the collection of these as the **ensemble-outcome archive**.

In the experiments in this thesis two different techniques are used to form ensembles:

1. **Inverse Noise** In this method, the initial condition ensemble is formed by adding random draws from the inverse of the distribution of observational noise on the true initial condition (observation) [31]. The inverse distribution is obtained by reflecting the noise distribution in the y axis. If the noise distribution has mean zero and is symmetric, the noise distribution and the inverse distribution are the same. With this Inverse Noise method, the initial states will not be consistent with the long-term model dynamics (e.g. they are not on the attractor, if there is one). This approach is applied in the experiments in chapters 4 and 6.

2. **Truncation** In this method, the initial condition ensemble is formed by adding random draws to the truncated (i.e. the number of decimal digits is limited) value of the observation. Truncation can be done by for example limiting the number of decimal digits. This technique is applied in the experiments in chapter 5.

Inside the PMS, more complex ensemble formation methods are justified [32]. These will not be considered in this thesis.

2.7 Summary

This chapter has defined terms used in this thesis. There is nothing new in this chapter beyond the presentation.

Chapter 3

Tools for Measuring Predictability

In this chapter, we describe background material on techniques used for computations and analyses in chapters 4, 5 and 6. Whilst there is no new derivation in this chapter, the distinction between two methods of computing the Ignorance [47, 88] and the Information Deficit [34] is new and as is the formulation in section 3.3.

The overarching theme of this thesis is predictability. In section 3.1 we give our definition of predictability and explain how it decays. In section 3.2 we introduce the concept of skill scores, give some examples and define some properties of them. One of the scores we define is the Ignorance score which is of particular interest in this thesis in terms of measuring the quality of the forecasting scheme. In section 3.3 the Climatological Ignorance [34], the Empirical Ignorance [34] and the Model Implied Ignorance [34] by method are introduced using two slightly differing approaches. In section 3.4, a diagnostic tool called Information Deficit [34] is introduced. In section 3.5, we describe the Relative Entropy, which we use later in the thesis to quantify the difference between two probability distributions. In section 3.6 we

describe procedures required to convert ensembles into probability forecast densities. A method called kernel dressing [19] is then defined and the concept of blending with climatology [19] is also introduced. Akaikes Information Criterion [10, 9] and cross-validation, two methods of model selection used in this thesis are then described.

3.1 Predictability

The main theme of this thesis is the predictability of certain real-valued observations from a dynamical system. Before proceeding, we require a definition of what exactly we mean by predictability. In fact, there is no universally accepted definition of predictability and its meaning may differ both within and between different disciplines. Many attempts have been made to define and quantify predictability. Quantifying the predictability of dynamical deterministic systems is a subject of statistical or dynamical studies [84, 99, 58], as well as philosophical investigation [39, 79]. Smith [100] suggests that what a mathematician may think of as predictability may well be very different from a physicist's conception of it. In this thesis, we follow Smith's mathematician's definition, which states that predictability can be understood as a property that allows the creation of a forecast distribution that is expected to be more informative than the final (climatological) distribution.

A number of different variables are forecast in this thesis. In chapter 4 and 5 the predictand[1] is the x variable of low-dimensional mathematical models such as the Henon Map, Ikeda Map and Logistic Map. In chapter 6 the target of the forecast is a computer generated observation of the model sea surface temperature anomaly in a specific region of Pacific in °C. In chapter 7 the predictand of our interest is radiation upwards in W/m^2, temperature in K and precipitation in kg/m^2.

[1] Predictand is a value to be predicted, which in this thesis is a real valued scalar quantity.

3.1.1 The decay of predictability

We explained in section 2.4 that a model can be used to predict the behaviour of a dynamical system. For a model to be useful in terms of prediction, it is expected that, at least at early lead times, forecasts generated using the model are able to outperform the climatological distribution. We have already defined the climatology in section 2.2. As the lead time is increased, the average skill of the forecasts is expected to decrease and eventually they become unable to outperform the climatology. We call the drop in skill with lead time **the decay of predictability** and say that predictability is lost when, on average, the forecasts perform no better than the climatology [103].

An important issue regarding predictability is how long the trajectory of the model remains close to the true trajectory of the system. Later in this chapter, measures of predictability such as the Ignorance and the Information Deficit are introduced. In chapter 5 we apply these measures to a low-dimensional mathematical model, while in chapter 6 they are applied to a deterministic numerical model of the coupled ocean and atmosphere, which is used to study El Niño events.

3.1.2 Limitations to predictability

There are several reasons that may limit the predictability of even a deterministic dynamical system. These are [78], most notably:

1. Initial condition uncertainty. In nonlinear systems, small initial errors can grow very quickly leading to larger errors later down the line [65].

2. Structural model inadequacy. A model is only an approximate representation of the laws of nature. When model inadequacy is present, the equations used to

describe the model differ from those of the system, e.g. a function is incorrectly defined.

3. Parameter error. The estimated parameter values may differ from the 'true' parameter values. Note that, in the IMS, arguably no 'true' parameter values exist.

For a stochastic dynamical system the random changes constitute a limit to the predictability. This has a similar effect to initial condition uncertainty which means that forecasts should be probabilistic in nature. In stochastic systems, even knowledge of the precise initial state and a perfect model does not allow precise point predictions. In a chaotic system it would. Operationally, the decay of information in a probability forecast of a stochastic system is evaluated in the same manner as probability forecast of a chaotic system with uncertain initial condition.

3.2 Skill Scores

Probabilistic forecasts are assessed using functional of the forecast distribution and the real valued outcome called skill scores. The discussion about what is being forecast in the thesis can be found in section 3.1. Scoring rules can be used to assess the relative performance of two forecasting systems. A skill score is a function of a probabilistic forecast and a corresponding outcome $S(p(y), Y)$, where Y is the outcome and $p(y)$ is the probability density. By convention, skill scores are defined to be negatively orientated, that is, the lower the score the better the forecast. It is not appropriate to evaluate the performance of a single probabilistic forecast. As a result, the mean score of many comparable (i.e. using the same forecasting system and lead time) forecasts and outcomes can be taken as a measure of forecast

performance.

A wide variety of scoring rules have been suggested over the years. These include, for example [102], the Root Mean Squared Error ($RMSE$) [36], the Naive Linear Score [37], the Proper Linear Score [56] and the Continuous Ranked Probability Score ($CRPS$) [35]. In order to choose an appropriate scoring rule, a number of useful properties can be taken into account. Thus a scoring rule is **proper** [112, 13] if it is optimised when the distribution from which the outcomes are drawn is issued as the forecast. A scoring rule is proper if the following inequality holds:

$$\int_{-\infty}^{\infty} S(p,y)q(y)dy \geq \int_{-\infty}^{\infty} S(q,y)q(y)dy, \tag{3.1}$$

where p is the model density and q is the system density.

Propriety is a desirable property of skill scores because it encourages a forecaster to issue their true belief as a forecast. Another property of skill scores is locality. A score is **local** [13] if it only takes into account the forecast density or probability at the outcome rather than the entire distribution. Further discussion of propriety and locality can be found in [22, 72, 56, 35, 36].

3.3 Ignorance

One particular scoring rule is the **ignorance score**. The ignorance score is the only scoring rule that is both proper and local [44, 68]. It is also often referred to as the logarithmic score and is defined by [47, 88] as

$$S(p,Y) = -\log_2(p(Y)) \tag{3.2}$$

where $p(Y)$ is the density or probability placed on the outcome Y by the forecast.

Due to its desirable properties, the ignorance score is the only score we use throughout this thesis. We are going to use this measure to assess a forecast in chapter 5, where we use a mathematical model - Logistic Map to model another more complex mathematical model - Quartic Map. Ignorance is also applied in chapter 6 to an oceanic-atmospheric model of El Niño events.

In this thesis we use three types of Ignorance: Climatological Ignorance, Empirical Ignorance and Model Implied Ignorance.

3.3.1 Climatological Ignorance

The ignorance score, as defined in equation 3.3, gives no indication of absolute skill. To be useful, a forecasting system should outperform the climatology. If this were not the case, a forecaster would be better off issuing the climatology as the forecast. We therefore define the ignorance of the climatology to be the zero skill level. This is denoted p_c [34].

There are two competing ways to normalise ignorance. We will call these IGN_{ClimE} and IGN_{ClimD}. IGN_{ClimD} is described in section 3.3.4. Following Bröcker and Smith [19], we define IGN_{ClimE} to be

$$IGN_{ClimE} = -\log_2(p_c(Y)), \tag{3.3}$$

The quantity IGN_{ClimE} corresponds to the ignorance of the actual outcome when using the climatology as the forecast.

In the discussion below we adopted method E for normalisation (see eq.3.3). The

reasons for this choice of method are made clear in section 5.2.4.

3.3.2 Empirical Ignorance

The Empirical Ignorance quantifies the skill of a probabilistic forecast. Computing the Empirical Ignorance requires the outcome to be known. In this thesis, following [34] we define the Empirical Ignorance to be the ignorance of a forecast relative to climatology as

$$IGN = -\log_2(p(Y)) - (-\log_2(p_c(Y))), \tag{3.4}$$

where Y is an outcome.

The **Mean Ignorance** of a set of n forecasts is given by

$$< IGN >= \frac{1}{n}\sum_{i=1}^{n}(-\log_2(p_i(Y_i)) - \frac{1}{n}\sum_{i=1}^{n}(-\log_2(p_c(Y_i))), \tag{3.5}$$

where $p_i(Y_i)$ is the probability placed on the i_{th} outcome by the i_{th} forecast.

3.3.3 Model Implied Ignorance

The Model Implied Ignorance is the expected value of the ignorance in the case where the outcome is a random draw from the forecast distribution itself [34]. The Model Implied Ignorance provides a quantitative idea of how much information one would expect to gain from having the forecast rather than climatology if the model

forecast distribution was perfect. The distribution of differences between Empirical Ignorance and Model Implied Ignorance gives an indication the inadequacy of the forecast.

Following [34], we define the Model Implied Ignorance of a forecast to be the expected ignorance, relative to climatology, that would be achieved were the outcome drawn from the forecast itself. Specifically

$$IGN_{MI} = -\int_{-\infty}^{\infty} p(y)\log_2\left(p(y)\right)dy - \Big(-\int_{-\infty}^{\infty} p(y)\log_2\left(p_c(y)\right)dy\Big), \tag{3.6}$$

where $p(y)$ is a forecast density and $p_c(y)$ is the climatology. If $p(y)$ approaches $p_c(y)$, the Model Implied Ignorance approaches zero.

3.3.4 Alternative normalisation

In section 3.3.1 we defined the zero skill level of the ignorance score to be the ignorance of the outcome when using the climatological distribution as a forecast. An alternative normalisation, defined by Du and Smith [34], is given by

$$IGN_{ClimD} = -\int_{-\infty}^{\infty} p_c(y) log_2 p_c(y) dy, \tag{3.7}$$

This method corresponds to the expected climatological ignorance rather than the climatological ignorance of the outcomes. We will refer to this approach as method D.

In this case, the Empirical Ignorance becomes

$$IGN = -\log_2(p(Y)) - IGN_{ClimD} \tag{3.8}$$

and the Model Implied Ignorance becomes

$$IGN_{MI} = -\int_{-\infty}^{\infty} p(y)\log_2(p(y))dy - IGN_{ClimD} \tag{3.9}$$

Two different ways of calculating the Empirical Ignorance and the Model Implied Ignorance are explored in chapter 5.2. In section 5.2.4, we argue that method E yields a more useful measure of the performance of a set of forecasts and hence method D is dropped. Thereafter, each time we refer to the ignorance score, the quantity IGN_{ClimE} (eq.3.3) has already been used for normalisation unless otherwise stated. Method E is then adopted and applied in section 5.3, and in the case of a real world forecasting model for El Niño predictions in sections 6.3 and 6.4.

3.4 Information Deficit

The concept of the Information Deficit was introduced by Du and Smith in 'Parameter estimation through ignorance' [34]. Following [34], we define the Information Deficit to be

$$Inf_{Def} = IGN - IGN_{MI}, \tag{3.10}$$

where IGN and IGN_{MI} are the Empirical Ignorance and the Model Implied Ignorance respectively of a given forecast. This definition differs slightly from that given by the authors of this paper who do not define the Empirical Ignorance and the Model Implied Ignorance in terms of the raw ignorance (i.e. not relative to climatology). Inasmuch as the raw ignorance alone is hard to interpret, we believe that our approach is more informative and can be considered an extension of the approach proposed by Du and Smith.

The Information Deficit can reveal inadequacies in the forecast system and quantify the closeness of the predictability of the current model to (its internal) perfection. If the Model Implied Ignorance is the same as the Empirical Ignorance, the Information Deficit is, by definition, zero. In a perfect forecast scenario, that is one in which the outcome is truly drawn from the forecast density, we expect, on average, for the Information Deficit to be zero. In general, if the forecast is imperfect, the Information Deficit is not expected to be zero. When the climatology is used as a forecast, the Information Deficit is expected to be zero. When forecasts are blended with climatology, the forecast density approaches the climatology. Therefore, the Information Deficit is expected to approach zero. In sections 5.2, 5.3, 6.3 and 6.4 we develop and demonstrate a novel approach to the use of the Information Deficit as a diagnostic tool.

3.5 Relative Entropy

Relative Entropy (RE) or Kullback-Leibler divergence [67] is a measure which can be used to quantify the difference between two probability distributions P and Q. In the discrete case, the RE from Q to P is given by

$$RE(P \parallel Q) = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)} \right), \tag{3.11}$$

where i is the number of possible values of P.

The RE can be interpreted as the information lost, in bits, when a true Q is approximated using some distribution P. The RE is always non-negative [83] and is equal to zero only when the two distributions are identical.

The RE can be used to measure the difference between two continuous distributions given a sample from each. This is done by defining N_b number of intervals, called bins, and counting how many members from each sample fall into each bin. This defines two discrete probability distributions P and Q.

Two ways of choosing the bins are considered in this thesis:

1. RE with equally likely bins. Each bin is defined such that P is a discrete uniform distribution. As a result, the width of each bin is allowed to vary and, if the two samples are the same, the probability of falling into each bin is $\frac{1}{N_b}$.

2. RE with equally spaced bins. Each bin is defined according to a fixed, identically spaced grid. This means that, usually, the distribution of P will vary

depending on i, that is the probability varies with each bin. Here we restrict application to one dimension where equally spaced bins are easily obtained by rank ordering.

Relative Entropy and Ignorance

The Relative Entropy is a measure of the difference between two distributions. For continuous distributions $p(x)$ and $q(x)$ equation 3.11 has the form:

$$RE(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) dx, [15] \tag{3.12}$$

Applying simple algebra to equation 3.12 yields:

$$\begin{aligned} RE(p \parallel q) &= \int_{-\infty}^{\infty} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) dx \\ &= \int_{-\infty}^{\infty} p(x) \log_2(p(x)) dx - \int_{-\infty}^{\infty} p(x) \log_2(q(x)) dx, \end{aligned} \tag{3.13}$$

In the particular case when we consider the RE of the model with respect to climatology, that is, setting $p(x) = p_m(x)$ and $q(x) = p_c(x)$ in equation 3.13, where $p_m(x)$ is the model distribution and $p_c(x)$ is the climatological distribution, equation 3.13 becomes:

$$RE(p_m \parallel p_c) = \int_{-\infty}^{\infty} p_m(x) \log_2(p_m(x)) dx - \int_{-\infty}^{\infty} p_m(x) \log_2(p_c(x)) dx, \tag{3.14}$$

Equation 3.14 quantifies the information lost (in bits) when the model distribution is approximated with the climatology.

From this we can note that:

$$RE(p_m \parallel p_c) = -IGN_{MI} \tag{3.15}$$

This is not surprising, since the Model Implied Ignorance can be interpreted as the expected gain in information (in bits) from having the model distribution as opposed to the climatology, assuming that the model distribution is perfect.

3.6 From Ensembles to Probability Forecasts

We introduced ensembles in chapter 2.6. Raw ensembles are inconvenient to interpret. It is often useful to convert ensembles into a probability density function which acts as a probabilistic forecast called a **forecast density**. In addition to being easier to interpret, forecast densities are arguably easier to evaluate than raw ensembles [111]. In this thesis, we use the performance of forecast densities as a method of quantifying predictability.

There are many possible approaches to converting ensembles into forecast densities. For example, a parametric distribution such as a Gaussian distribution could be fitted to the ensemble [89, 113, 28]. For nonlinear models, in which the ensemble is unlikely to follow a known parametric distribution, a more flexible approach, however, is to use kernel density estimation [26, 74, 50] which is defined below.

3.6.1 Kernel density estimation

Kernel density estimation is a nonparametric method to estimate the probability density function from a finite sample [80, 91]. Given a sample $x = x_1, .., x_n$ from some unknown distribution f a kernel density estimate of f is defined by:

$$p(x : X, \sigma) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{\sigma}\Big), \tag{3.16}$$

where K is the Gaussian kernel as defined in eq.3.18. The shape of the density is determined by the kernel width σ . Many different approaches to the selection of the bandwidth parameter exist which usually make some assumption about the underlying distribution of the sample [57, 17].

3.6.2 Kernel dressing

It has been argued [19] that constructing forecast densities by fitting parametric distributions or using a nonparametric approach such as kernel density estimation have a serious shortcoming. Whilst both of these approaches aim to estimate the distribution of the ensemble, in fact, we wish to attempt to estimate the distribution of the outcome [45, 19]. Following [19] we now describe an approach to constructing forecast densities called kernel dressing. Forecasts formed using kernel dressing take the form

$$p(y|x, a, o, \sigma) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\Big(\frac{y - ax_i - \upsilon}{\sigma}\Big), \tag{3.17}$$

where K is a kernel function which, we will take to be Gaussian of the form

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \tag{3.18}$$

Here x is the predictand, $x_1, .., x_n$ is an ensemble, a is a scaling parameter, υ is an offset parameter and σ is the kernel width, that is, the standard deviation of the Gaussian kernel. The value of σ, which affects the shape of the distribution, is chosen as that which optimises the mean performance, with respect to ignorance score, over some archive of ensembles and corresponding outcomes. A new ensemble is then dressed using the optimised kernel width σ.

Note that Binter [14] argues that kernel density is distinct from kernel density estimation. The kernel density estimation approach aims to find the distribution from which the sample was drawn. In kernel dressing, the underlying distribution of the ensemble and the outcome are different, so the purpose of this procedure is to find the most useful forecast with the respect to the scoring rule, not the distribution for the next ensemble member.

3.6.3 Blending with climatology

As described in section 3.1, the climatology provides a benchmark for the performance of a set of forecasts. The concept of blending [19] ensures that a set of forecasts is never expected to perform worse, on average, than the climatology. Blending the climatological distribution of the system $p_c(x)$ with a distribution derived from the model $p_m(x)$ yields a probability forecast

$$p(x) = \alpha p_m(x) + (1 - \alpha) p_c(x), \tag{3.19}$$

where $\alpha \in [0, 1]$ is called a **blending parameter**. When used in conjunction with kernel dressing, the blending parameter is optimised simultaneously with σ over the ensemble-outcome archive.

The value of α provides an indication of the usefulness of the model-based forecasts. When α is significantly greater than zero, the information in the model is useful because it provides better forecast skill than purely using the climatology as the forecast. At longer lead times, as the information contained in the ensemble decays, we expect the blending parameter to approach zero[2]. In that case, the blended forecast and the climatology coincide and the model forecasts no longer contain useful information. Unless stated otherwise, all forecasts in this thesis have been blended with a climatology.

3.6.4 Model selection techniques

Akaike's Information Criterion (AIC) [9, 10] is a way of determining whether additional parameters are justified in a model structure. Following [24] Akaike's Information Criterion (AIC) is given by

$$AIC = 2r - 2\ln(L), \tag{3.20}$$

where r is the number of free parameters fitted for the model to be assessed and L is the maximised likelihood.

The model with the lowest AIC value is preferred. AIC penalises for adding additional parameters (the first term in equation 3.20) and rewards models with the best

[2]Note this is not necessarily true in the perfect model scenario in which, at longer lead times, the ensemble can be considered to be drawn from the system climatology. In this case, the optimal value of α will depend on the size of the ensemble.

fit as measured by the log-likelihood function (the second term in equation 3.20).

AIC is used in section 5.2.2 in order to determine when the model contribution should be set to 0.

3.6.5 Cross-validation

An alternative approach to model selection is to use cross-validation. In cross-validation, parameters are optimised over subgroups of some set and the performance tested over others. Parameters of the model are robust if they perform well out-of-sample. Throughout this thesis we use a leave-one-out cross-validation which is defined below.

Leave-one-out cross-validation

In leave-one-out cross-validation, the parameters are optimised over all but one data point and the performance evaluated over the remaining point [54]. This procedure is repeated by leaving out testing on each of the points exactly once. The method can be computationally expensive, as it requires to carry out optimisation many times. In this thesis the climatological distribution is formed using kernel density estimation and the kernel width is estimated using leave-one-out cross-validation minimising the mean ignorance score.

3.7 Conclusions

This chapter has introduced a number of concepts and measures for quantifying predictability that are developed and deployed throughout this thesis. Methods of converting ensembles into probability forecasts are described and methods of model selection are defined. While no new theory is introduced in this chapter, its novelty

goes beyond only presentation. A novel distinction has been made between two different approaches of normalising the Ignorance score and the Information Deficit. One method of normalisation is concluded to be superior to the other and it is suggested that the alternative approach is dropped.

Chapter 4

Distinguishing model inadequacy from chaos in nonlinear simulation

This chapter addresses the question of distinguishing model inadequacy from chaos in a nonlinear simulation. Sensitivity to initial conditions complicates the forecasting of chaotic dynamical systems, even when the model is a perfect representation of the system generating the observations [103]. Structural model inadequacy is a distinct source of forecast failures, although it is sometimes questionably used as in an example [77] to be due to uncertainty in the initial condition. The resource allocation in treatment of these two limitations of predictability is very different. To reduce the impact of model inadequacy, one can improve the existing model, build an independent model or combine a collection of models [82, 30, 60]. To cope with uncertainty in the initial condition, we could lower the noise level, build a larger ensemble [73, 66] or use some data assimilation technique [32, 33].

In this chapter we make a clear distinction between model inadequacy and chaos

and study those differences.

This chapter is structured as follows. In section 4.1, the effects of model inadequacy and chaos are visualised using the extremely similar nonlinear dynamics of the single and double precision Henon Maps. In section 4.2, an experiment to distinguish model inadequacy and chaos is carried out on the Henon Map as a model of its more mathematically complicated version - the Senon Map [93]. Here, we investigate the evolution of the Relative Entropy of the same initial distributions of ensembles under the system and the model. Weather-like and climate-like analogies are explained. Next, the System-Model pair considered in section 4.3 are the Ikeda Map and truncated Ikeda Map. Again the distinction between model inadequacy and chaos is made and the impact of the location on the system attractor of the two initial conditions is illustrated. Finally, section 4.4 presents conclusions for this chapter.

4.1 Illustration with Henon Map: single and double precision

4.1.1 Divergence of specific Initial Conditions

First we demonstrate the concept of sensitive dependence on the initial condition. Small errors such as noise or computer roundoff can change the final solution significantly. As an illustration of this, Fig.4.1 shows the results of a computer experiment on the two-dimensional Henon Map, which is defined in Appendix A.1.

In each chart in Fig.4.1 the green points show the position of the 32nd to 42nd states from two trajectories initialised with two slightly different initial conditions (black

dots) with the double precision Henon Map. The blue dots represent randomly drawn points from the attractor. In the top chart, the initial condition is a transient state far from the attractor whilst in the lower chart, the initial condition is taken from the attractor. The distance between two green points (straight black line) depends on where on the chaotic system the initial condition is located.

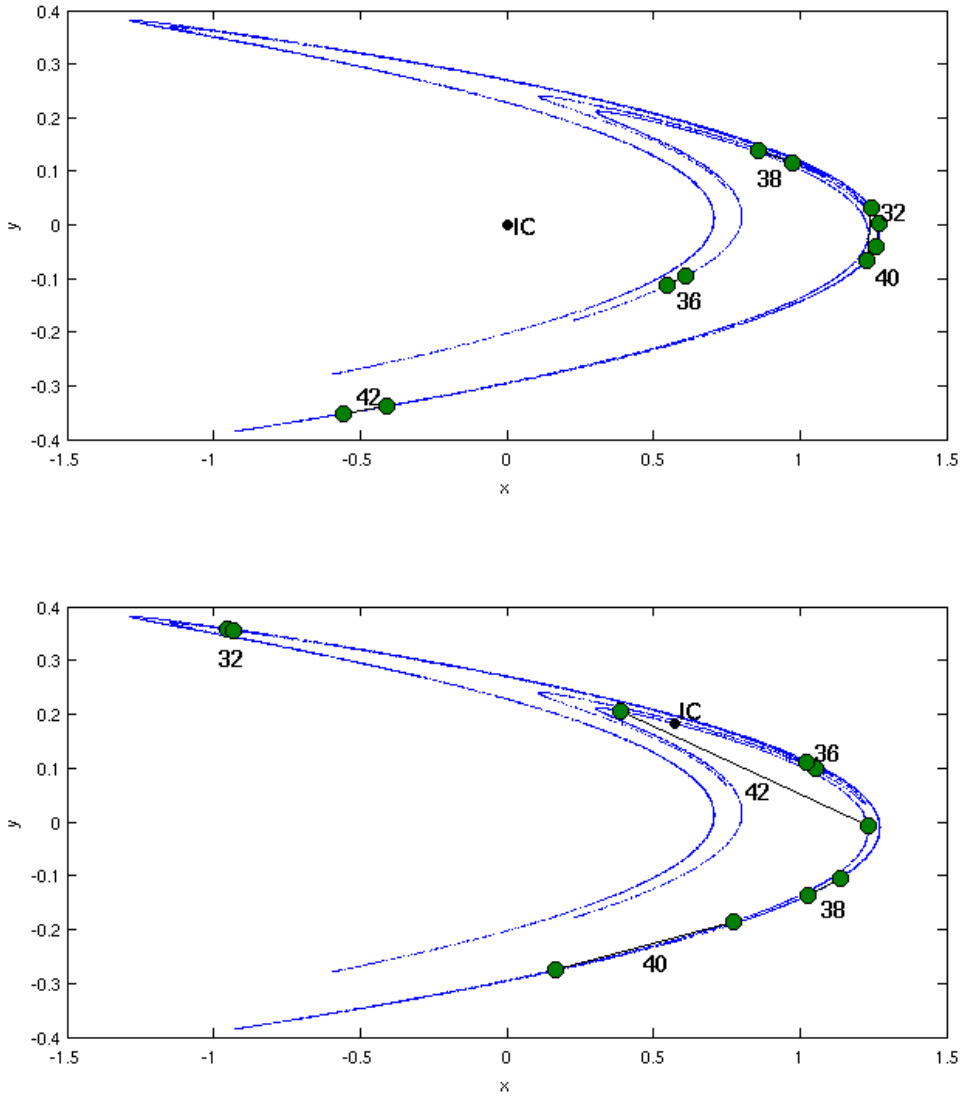


Figure 4.1: The two panels show two trajectories with slightly different initial conditions iterated under the same dynamics. The separation between these initial conditions is $\varepsilon = (2^{-10}, 2^{-10})$. After 32 to 40 iterations the trajectories look very different. The distance between a pair of points (black line), depends on the location of the initial condition. For example distance at lead time 40 is bigger in the bottom chart than in the top one.

The effect of model inadequacy is exemplified in Fig.4.2. Here, the same initial

condition was evolved with two different sets of dynamics: the Henon Map in single precision and in double precision. The yellow and red circles show the 32nd to 42nd iterates of the single and double precision trajectories respectively. The IC has a large effect on the distance between the states of the two trajectories.

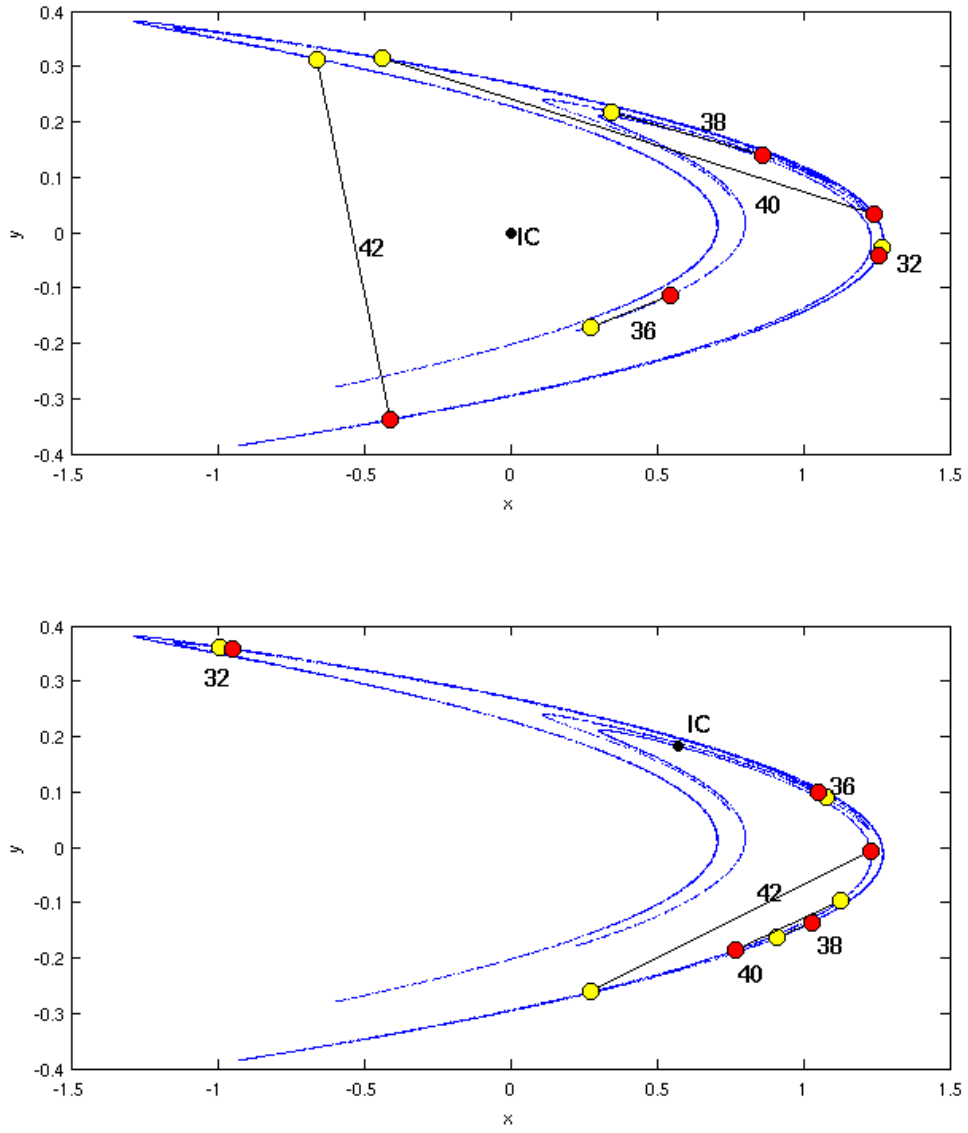


Figure 4.2: The same Initial Condition (IC) was evolved under two different dynamics: single (yellow) or double (red) precision Henon Maps. Somewhere between 32 and 40 iterations later both trajectories originating from the same initial condition are far apart. The distance between points (black line), depends on where on the chaotic system the initial condition is built.

4.1.2 Divergence averaged over the attractor

We have shown that the distance between two trajectories varies with the location from which the initial condition is drawn. Now, we consider a larger number of initial conditions to demonstrate how, on average, this distance changes with time.

Fig.4.3 shows the base 2 log of mean Euclidean distance d between 8196 pairs of points generated using single and double precision Henon Maps. This is shown as a function of time in the top left chart of Fig.4.3. The top right panel illustrates the same, but with both trajectories generated using double precision with the initial separation of ε=2^{-24} units. For clarity, the lower panel shows the difference between the two. The average distance between trajectories in both cases is growing with time, from 2^{-24} at early lead times to around 2^{-1} at longer lead times.

These results demonstrate that model inadequacy and uncertainty in the initial condition are two different limitations in forecasting. This distinction is important, both in terms of model development and in terms of resource allocation, for improving observations in operational forecasting.

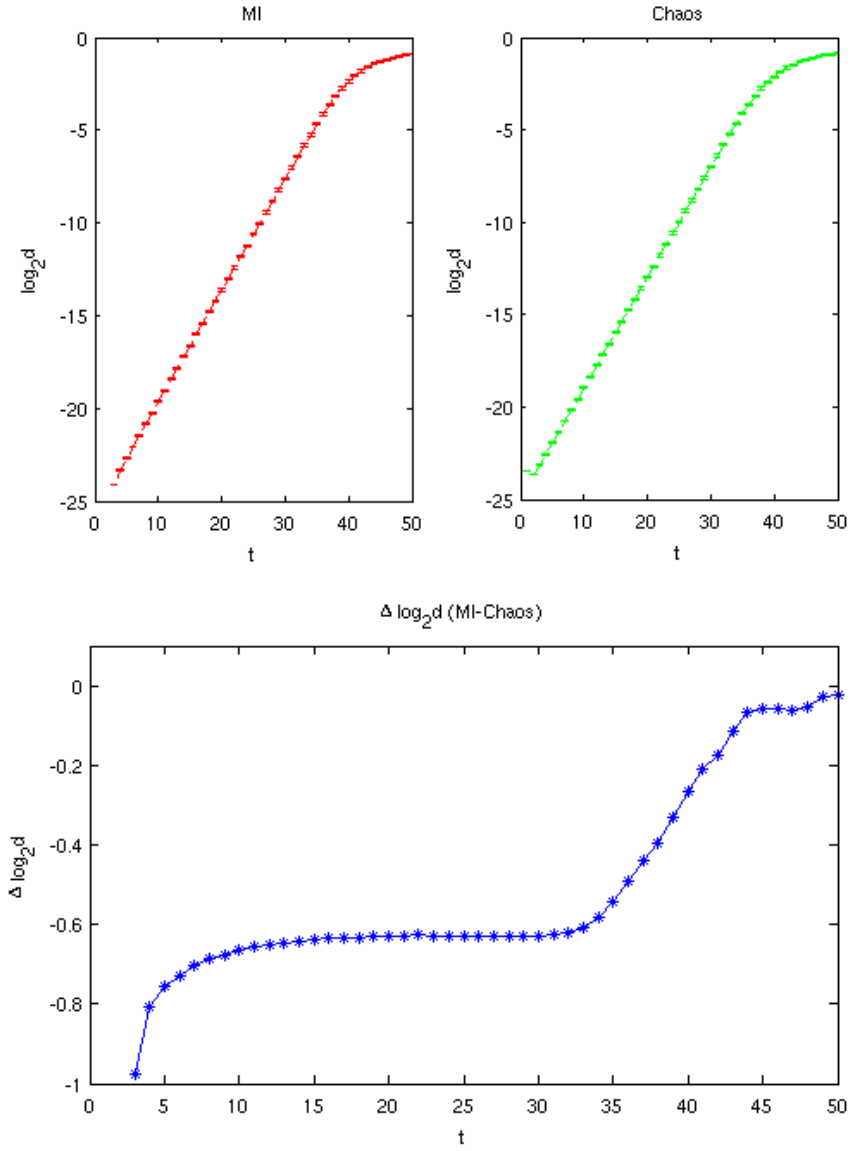


Figure 4.3: Mean Euclidean distance between trajectories originating from many distinctly located initial conditions of a pair of points as a function of lead time in a situation of model inadequacy (top left) and chaos (top right) as illustrated in Figs.4.1 and 4.2. Bars along the green and the red curves represent 95% re-sampling intervals of the mean. As shown in the top charts, the divergence of the trajectories increases with time in a different way. The bottom chart illustrates the difference between the two charts at the top. These results demonstrate that model inadequacy and chaos are two different sources of forecast failures and should be distinguished when considering future resource allocation.

4.2 Illustration with Henon Map and Senon Map

The previous section 4.1 illustrated the effect of model inadequacy and initial condition uncertainty using Henon Map. This time, the model is again defined as the double precision Henon Map. The system, however, is the Senon Map introduced by Smith in 2002 [93] and described in Appendix A.2. The attractors of these two simple nonlinear two-dimensional chaotic maps appear remarkably similar, as shown in Fig.A.1. In this section we carry out an experiment which aims to distinguish model inadequacy from chaos.

4.2.1 Experimental design

The scheme of the experiment 4.A is illustrated in Fig.4.4, while the details are given in a table 4.1.

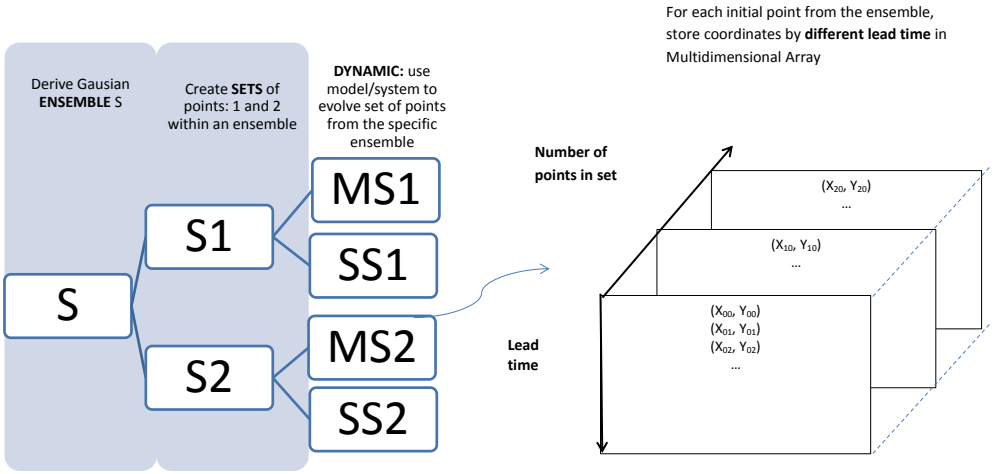


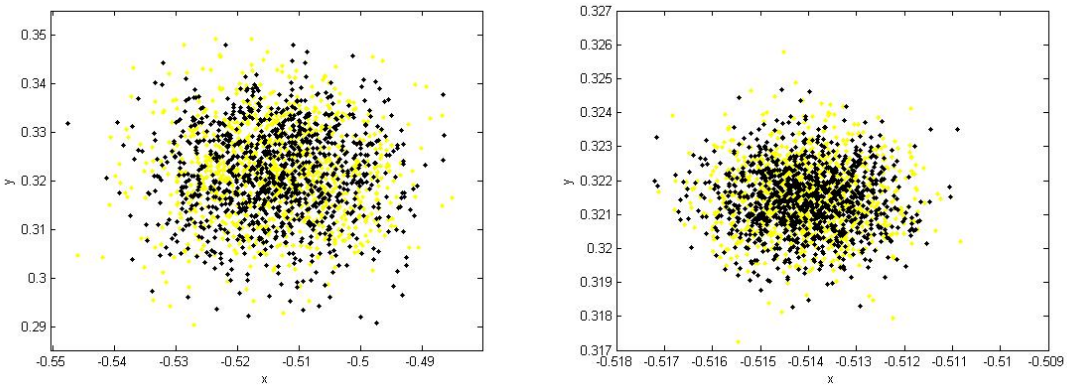
Figure 4.4: Processes in Experiment 4.A.

System	Senon Map
Model	Henon Map
A(x_0,y_0)	(-0.514, 0.3215)
n_{ens}	1
n_{sets}	2
m	4096
ts	64
S_{d_1}	0.01
S_{d_2}	0.001
Number of bins (RE)	64

Table 4.1: Experimental Design 4.A

In this experiment we derive an ensemble S, which is a Gaussian ball of points built around a point A randomly drawn from the attractor. Different sizes of the

Gaussian ball are considered with three different standard deviations (S_d): 0.01 and 0.001. Each ensemble consists of m members and is divided into two sets '1 and 2', of randomly chosen points as shown in Fig.4.5. Further, a model and a system are used to evolve each set of initial points (S1 and S2) from each ensemble and their positions are stored for each of the first 64 steps. In this experiment we only analyse x variable. We define SS1 as set '1' evolved on the system, SS2 is set '2' evolved on the system, MS1 is set '1' evolved on the model and MS2 is set '2' evolved on the model.



(a) $S_{d_1} = 0.01$ (b) $S_{d_2} = 0.001$

Figure 4.5: Sets of points '1' (yellow) and '2' (black) within a Gaussian ball for each standard deviation (S_d). Note the charts have different scales on their X and Y axes.

To capture the model inadequacy, we analyse how one set of points behaves under the model and the system at different time steps. In this case we will compare (SS1 to MS1) or (SS2 to MS2) for each time step. To capture the effects of chaos we analyse how set '1' and set '2' act under the same dynamics. In this case we will compare (SS1 to SS2) or (MS1 to MS2) for each time step.

4.2.2 Results

The question of interest in this experiment is whether the points evolved in time can be distinguished in a way that allows us to identify which mathematical structure they are being evolved under.

The relative entropy, defined in section 3.5 is used to measure the difference between the sets. Fig.4.6 illustrates relative entropy calculated using 64 equally likely bins, plotted at the first 16 steps (left panel) and 64 steps (right panel). The System-Model series (yellow, black lines) show how each set of initial conditions behave under the system and the model. This allows us to capture the effect of model inadequacy. Two slightly different sets of points, evolved using the same dynamics of the system or model, are represented using red and blue lines respectively. These illustrate the initial condition uncertainty. The lower panel shows a zoom of the red and blue lines. Fig.4.6 is the same as Fig.4.7, but concerns a smaller noise level ($S_d = 0.001$).

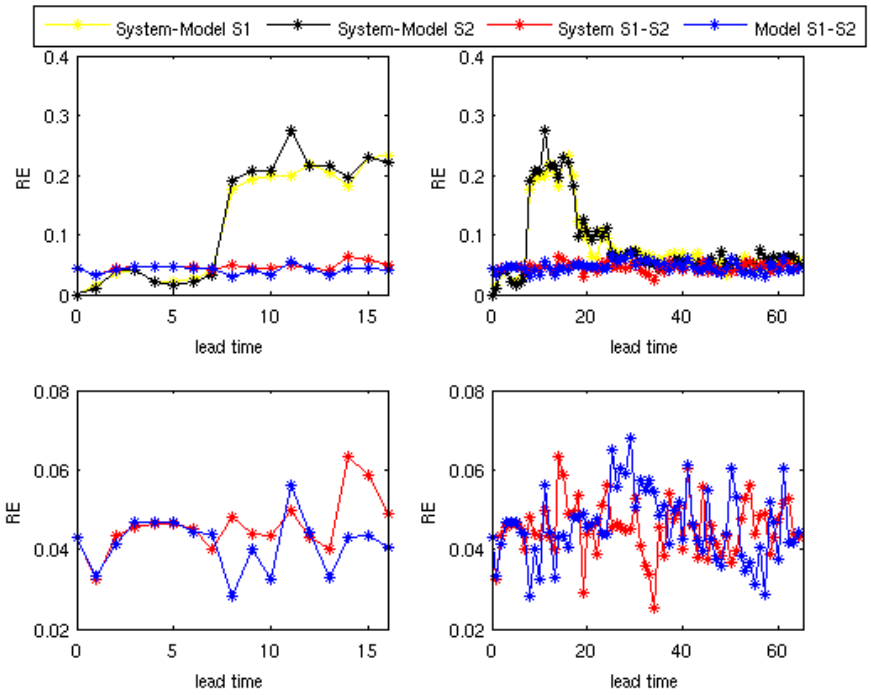


Figure 4.6: *RE* as a function of lead time for the first 16 time steps (left panel) and 64 time steps (right panel) in Experiment 4.A, for $S_d = 0.01$. The yellow and black lines show how one set of initial conditions behave differently under the different dynamics of the system and the model. The *RE* of S1 and S2 developed using the dynamics of the system or model is illustrated with red or blue lines respectively.

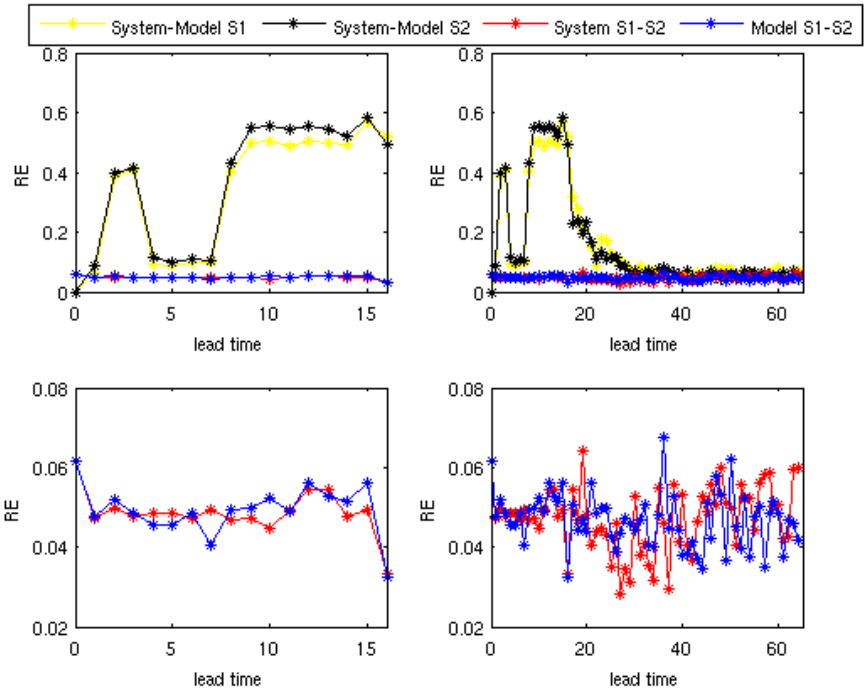


Figure 4.7: This figure is the same as Fig.4.6, but concerns smaller noise level ($S_d = 0.001$).

When the two dynamics are different, the RE starts at zero, grows relatively large and then converges back close to zero again. This is because the ensembles diverge over time and then approach their respective climatologies.

When the dynamics are the same but the sets of initial conditions are different, the ensembles do not diverge as much as when the dynamics are different.

When the noise level is lower, the ensembles take longer to approach their respective climatologies and hence the RE is higher for longer.

4.2.3 How does this relate to a weather-like climate-like context?

One way to think about predictability in nonlinear systems is to think about weather forecasting.

- In the short range (steps 1-7) the model is useful and reflects reality. The behaviour of the actual weather and the ensemble are hard to distinguish.
- In the medium range (steps: 8-20, a few days ahead in terms of weather forecast) the model cannot capture the true dynamics. The relative entropy, gives a very bad score for that.
- In the long run (steps 20+) the RE is smaller but the model is not likely to be any more useful than the system climatology, which in weather forecasting, is usually attainable from past observations. At this stage, the model is no longer likely to be informative.

4.3 Illustration with Ikeda Map

Next, we carry out an experiment on the Ikeda Map, which is structurally more complex than the Henon Map. In this experiment we forecast x variable. The details of Experiment 4.B are in Table 4.2, and the System-Model pair used here is presented in Appendix A.3. Similarly to the experiments described in the previous sections of this chapter, we first evolve a pair of the same initial conditions using two slightly different dynamics (that is the Ikeda Map and the Truncated Ikeda Map) to demonstrate model inadequacy. To demonstrate chaos, two initial conditions

separated by a very small distance ε are iterated forward using the same model dynamics (that is the Truncated Ikeda Map). Then we evolve a pair of initial conditions (for ts time steps) and measure the Euclidean distance between many (n_{ic}) pairs of points evolved at each time step. The threshold of our interest is time step when these points separate by 1.5 units.

System	Ikeda Map
Model	Truncated Ikeda Map
n_{ic}	8192
ts	50

Table 4.2: Experimental Design 4.B

Given that model uncertainty and initial separation are arbitrary a direct comparison between the time until the threshold is exceeded for model error and chaos is not possible. We note that the separation in model inadequacy case can be measured from time step one, at the initial time step there is no separation. To make analysis between model inadequacy and chaos directly comparable, the separation ε between two initial conditions at time step one in model inadequacy case is taken to be the same as separation between initial conditions at time step zero in chaos case.

In the upper panel of Fig.4.8, 8192 points from the model attractor are shown coloured according to the time taken for the model and system trajectories initiated at each location shown on the Ikeda Map to exceed 1.5. The lower panel shows the same, but for a case in which two trajectories are initialised with the same model dynamics using the point shown and another point with ε added to coordinates of the initial point. The threshold used is arbitrary, selected to give an indication of when two points are clearly far apart. Both diagrams relating to (a) model inadequacy and (b) chaos yield different results, which suggests that discussed concepts are different. The median in diagram (a) is 20, while in case (b) it is 6, thus a greenish

and redish colour is respectively dominant. Also, pattern on the attractors shows how the location of the initial conditions affects the time taken for the trajectories to diverge.

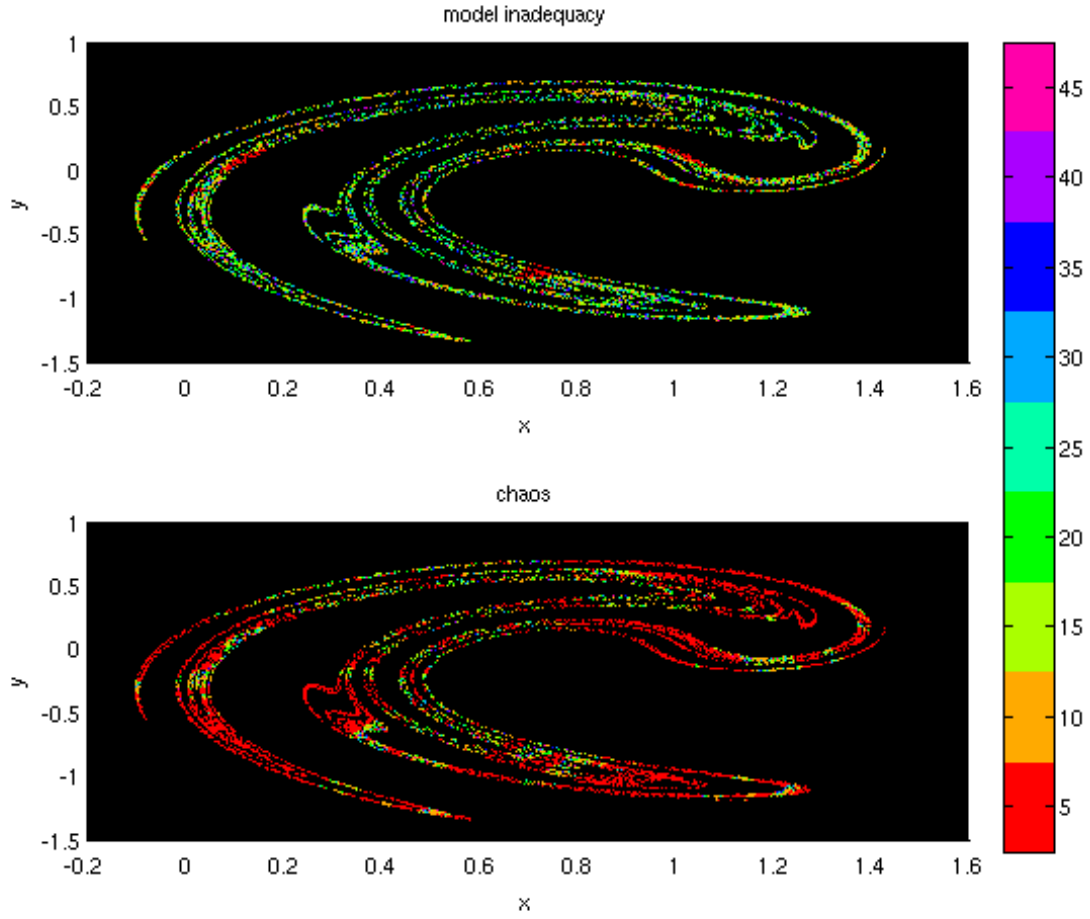


Figure 4.8: Chart (a) illustrates model inadequacy, chart (b) chaos in Experiment 4.B. For the initial conditions represented by dots shown on the Ikeda attractor, the colour indicates time steps until the Euclidean distance between trajectories originated from this location is greater than or equal to 1.5 units. The time taken to reach this threshold varies depending on the position of the initial condition on the attractor, thus we observe different patterns in each chart. The effect of model inadequacy and chaos is different, which is highlighted by the different overall pattern of colours of both diagrams.

4.4 Conclusions

In this chapter, we demonstrated that model inadequacy and chaos represent two different limits to predictability in nonlinear chaotic systems. This clear distinction has been studied using the Henon Map as shown in Fig.4.2 and the Ikeda Map in Fig.4.8. It has been demonstrated that these results vary by the location of the initial condition ensemble.

Other new aspects of this chapter are: an investigation of decay of predictability and distinguishing the effects on relative entropy of model inadequacy and chaos as shown in Fig.4.6. While Judd and Smith [95, 96] looked at the separation of points they did not analyse how decay of predictability varies the way it has been done in this chapter.

Chapter 5

The Decay of Predictability

The main focus of this chapter is to explore and expand the novel use of a measure called the Information Deficit in forecast evaluation. Here we also apply other predictability measures introduced in section 3.3, namely the Empirical Ignorance, and the Model Implied Ignorance, to demonstrate the decay of information in the context of mathematical low-dimensional system-model pairs.

In section 5.1, the model and the system are introduced and examined. Here, we investigate the difference between the natural measure of the Logistic Map and the Quartic Map as the magnitude of one of the parameters of the Quartic Map is changed. This is the first time the Quartic Map has been explored in this way. In section 5.2, we apply predictability measures defined in chapter 3 and examine how they decay in time. Here, two alternative methods of normalisation are considered (method D and method E). Several reasons are given to favour method E over method D. After this section, method D is not used again in this thesis. In section 5.3, the experiment is repeated with alternative versions of the system. Conclusions

for this chapter are presented in section 5.4.

5.1 Logistic and Quartic Maps

5.1.1 Model and System

Throughout this section, we define the model to be the **Logistic Map**. The Logistic Map is defined as

$$x_{t+1} = rx_t(1 - x_t), \tag{5.1}$$

where x is the model state and $x_t \in (0, 1)\forall_t$ and r is a parameter on the interval [0,4].

We define the system to be the **Quartic Map**, which is given by

$$\tilde{x}_{t+1} = (1 - \tilde{\mu})\tilde{r}\tilde{x}_t(1 - \tilde{x}_t) + \tilde{\mu}\frac{4}{5}[\tilde{r}\tilde{x}_t(1 - 2\tilde{x}_t^2 + \tilde{x}_t^3)], \tag{5.2}$$

where $\tilde{x}$ is the system state, and $\mu \in$ [0,1] and $r \in$ [0,4] are parameters. This was described in 'Laplaces Demon and the Adventures of His Apprentices' by Frigg, Bradley, Du, and Smith, 2014 [38].

When $\mu = 0$ equation 5.2 is equivalent to equation 5.1 and so the model is perfect. When $\mu = 1$, equation 5.2 becomes

$$\tilde{x}_{t+1} = \frac{4}{5}[\tilde{r}\tilde{x}_t(1 - 2\tilde{x}_t^2 + \tilde{x}_t^3)], \tag{5.3}$$

which we refer to as Full Quartic.

Model	μ	r	Equation
Full Quartic	1	4	5.3
Quartic	(0,1)	4	5.2
Pure Logistic	0	4	5.1

Table 5.1: These three maps will be repeatedly used in the chapter: Pure Logistic, Full Quartic and the map that mixes both together, which is called Quartic.

To contrast the case in which μ=0, that is the Pure Logistic Map, and that in which μ=1, we compare the invariant measure of each. To avoid making assumptions about the shape of each of the invariant measures, we take a simple approach in which the domain is divided into 64 equally spaced bins and the proportion of random draws that fall into each is compared. To get random draws from the invariant measure we use the following approach: take 32 random initial conditions uniformly distributed on (0,1), iterate them for 128 steps to disregard transient states, continue for 1024 steps recording the states of x. This is done 4 times so that we have 4096 states from each invariant measure in total. We refer to this experimental design as 5.A, the details of which are listed in Table 5.2.

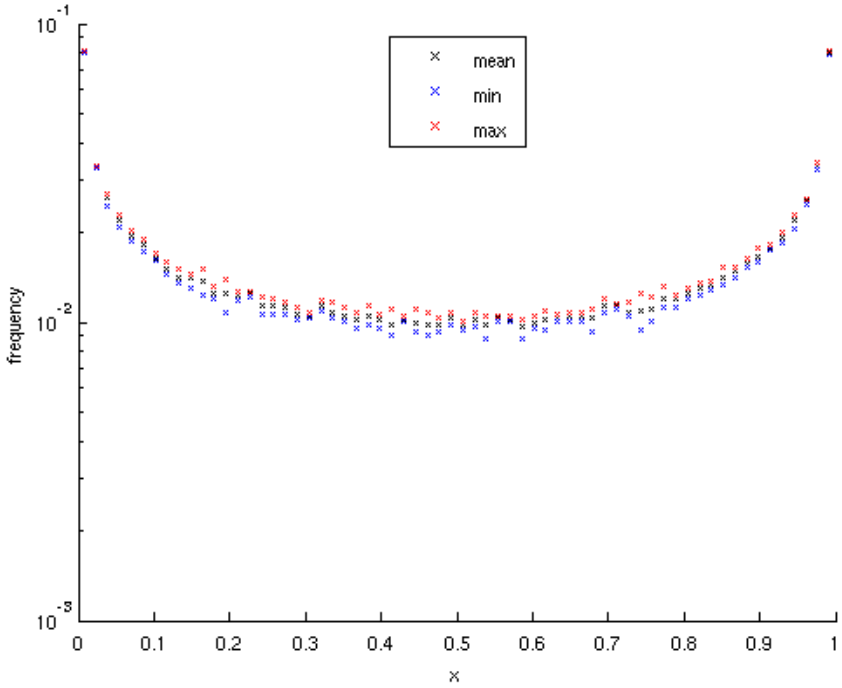
System	Quartic Map
r	4
Number of bins (distribution)	64
Number of runs	4
IC	U(0,1)
n_{ic} system	32
ts	1024
$TRAN$	128

Table 5.2: Experimental Design 5.A

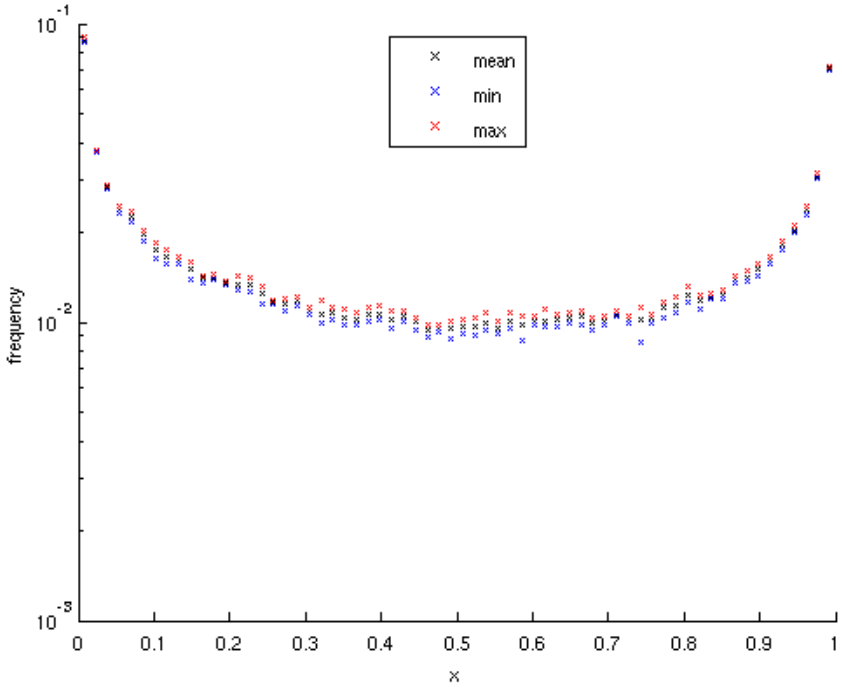
In Fig.5.1, for each bin, we show the minimum, maximum and mean proportion of

states falling into them over 4 different runs. This is done for a) $\mu = 0$ (model is Pure Logistic)[1] and b) $\mu = 1$ (system is Full Quartic). Here, we notice that both distributions have more probability around zero and one in both cases. However, it can be seen that each distribution is slightly different. For example, the proportions falling in the first bin are noticeably different.

[1]Note that the natural measure for the Logistic Map with r=4 is analytically known. It is the beta distribution with parameters a=0.5 and b=0.5. This is discussed in Appendix B.



(a) $\mu = 0$



(b) $\mu = 1$

Figure 5.1: The minimum (blue), maximum (red) and the mean (black) proportion of points falling into each bin over 4 sets of draws for a) $\mu = 0$ and b) $\mu = 1$ under the experiment 5.A. Both distributions have greater probability near x=0 and x=1. However the distributions are not the same. For example, there is a contrast between the min, mean and max in the first bin at the lower edge of the range of x.

5.1.2 Relative Entropy (RE) between the model and the system

Relative Entropy was defined in equation 3.11. In this section, we investigate the difference between the natural measures of the model and the system as the structural difference becomes smaller, that is as μ from equation 5.2 approaches 0. The question of interest here is: what is the smallest μ for which the distributions of the model and the system are distinguishable?

To answer this question, we compare the model and the system with different values of μ when $r = 4$. Details of the experiment, which we call experiment 5.B, are listed in Table 5.3.

System	Quartic Map
Model	Logistic Map
r	4
Number of bins (RE)	8
IC	U(0,1)
n_{ic} model	2048
n_{ic} system	1024
ts	1024
$TRAN$	128
Number of runs	4
μ	0; 2^n, where n= $\{-20, -19, -18, ..., 0\}$

Table 5.3: Experimental Design 5.B

The charts in Fig.5.2 show the RE, calculated using 8 equally likely bins in the system (see section 3.5 for details about the equally likely grid), between the invariant measure of the Model (randomly drawn ICs evolved using Logistic Map) and of the System (randomly drawn ICs evolved using Quartic Map) for $\mu = 2^n$, where $n = \{-20, -19, -18, ..., 0\}$. This was repeated 4 times and therefore results for run1, run2, run3 and run4 are shown on each chart. Chart 5.2 a) and b) show

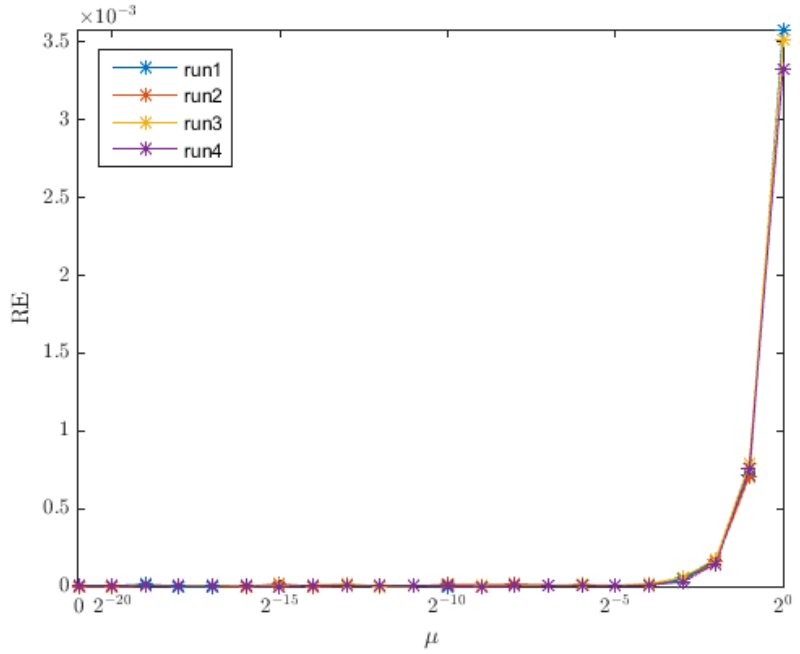
the *RE* between the model and the system for different ranges of μ. The dashed lines illustrate the range of the *RE* between the model and itself for different sets of model realisations from 4 different runs. This demonstrates the statistical variation of *RE* when system and model equations are the same (when μ=0).

The conclusions from this subsection are:

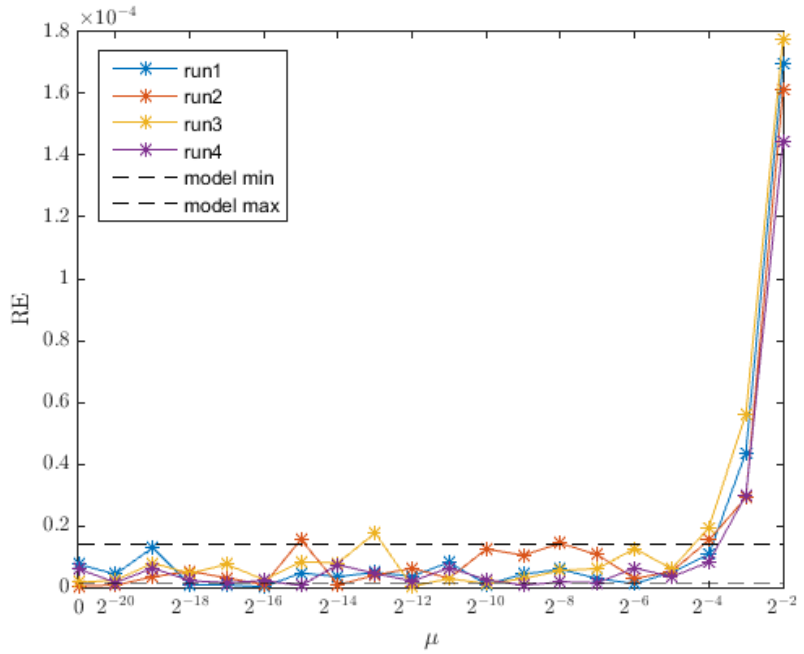
1. When $\mu \in (2^{-4},2^{0})$, the *RE* starts to increase. As expected, the *RE* is largest when $\mu = 1$ (or 2^{0}), because the system is the most different from the model.

2. When $\mu \leq 2^{-4}$ there is no obvious change in behaviour from the perfect model: the *RE* stays within the same range as the *RE* between the model and itself.

3. Points 1 and 2 above suggest that the system looks different from the model roughly when $\mu > 2^{-4}$ and similar to the model when $\mu \leq 2^{-4}$.



(a)



(b)

Figure 5.2: The RE between the model and the system for different values of μ in experiment 5.B. Note the different scale between chart (a) and (b). Each solid line corresponds to a different model run. The RE is largest when $\mu = 1$, which is when when the system is fully quartic. For $\mu \in (2^{-4},2^{0})$ RE increases quickly, for $\leq 2^{-4}$ it stays in a similar range. The dashed lines show the range of RE between model and itself for different model realisations. For $\mu > 2^{-4}$ there is clear evidence that the distributions of the model and the system are different.

System	Quartic Map
IC	U(0,1)
r	U(3.5,4)
n_{ic}	128
n_r	1024
ts	512
$TRAN$	128
μ_1	0
μ_2	0.1
μ_3	1

Table 5.4: Experimental Design 5.C

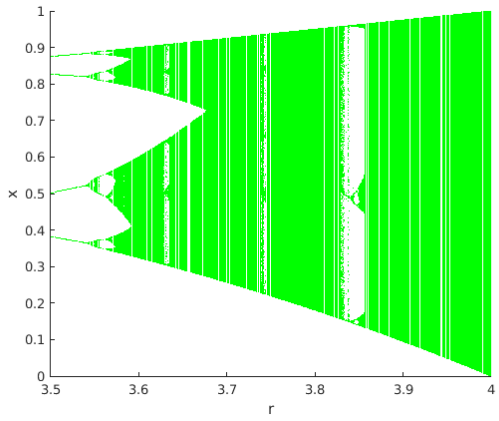
5.1.3 Bifurcation Diagram of the model and different versions of the system

In this section we illustrate additional properties of the model and the system. We examine the similarity of the behaviour of the model and the system using bifurcation diagrams. A bifurcation diagram shows the parameter r against values of x on the system attractor with that value of r for different values of μ.

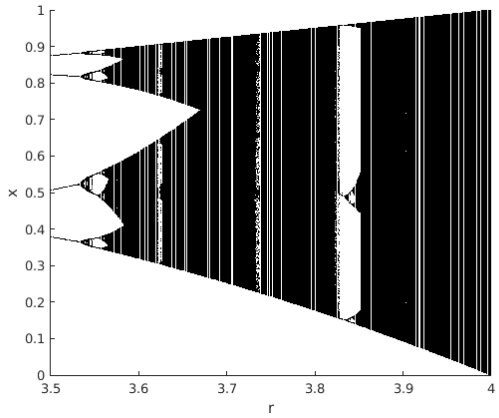
To generate the bifurcation diagrams, we drew 1024 random values of r from U (3.5,4) and drew 128 random ICs which were then run forward for 128 time steps (to ensure the states are not transient) and then iterated for another 512 time steps for 3 values of μ. We refer to this experimental design as 5.C. The details are listed in Table 5.4.

Fig.5.3 shows the bifurcation diagram for a) $\mu = 0$ (Logistic Map), b) $\mu = 0.1$ and c) $\mu = 1$ (Quartic Map). Here, periodic behaviour (when values of x oscillate around for example 3 or 6 values) starts at different values of r for each case. The period-3 window is the most prominent [81]. It happens around r=3.83, where out of the chaos, a stable 3-cycle appears. Here, this transition is shifted to the left between

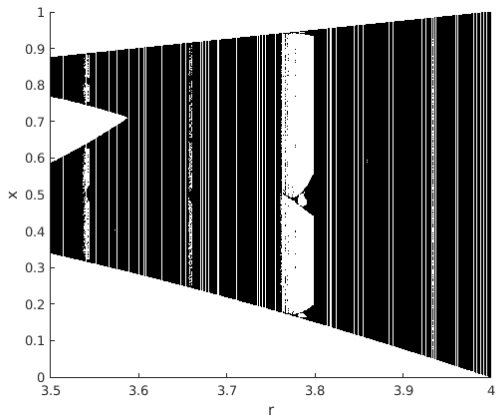
the top diagram and the bottom one. The period doubling cascade [81] is complete before r=3.5 for $\mu = 1$, unlike for $\mu = 0$ and $\mu = 0.1$, where this occurs in the range shown. Here, we conclude that the system with $\mu = 0.1$ is very similar to the model ($\mu = 0$), but when $\mu = 1$, the behaviour is very different.



(a) $\mu = 0$



(b) $\mu = 0.1$



(c) $\mu = 1$

Figure 5.3: Bifurcation Diagrams for experiment 5.C, where a) $\mu = 0$ (shown in green), b) $\mu = 0.1$ and c) $\mu = 1$ (system is Full Quartic). Qualitative behaviour did not change between diagrams a) and b). Diagram c) shows different behaviour. For example, with $\mu = 1$ the period doubling cascade is complete before r=3.5. Also, the period-3 window (near r=3.83) shifts left from a) to c).

5.2 Demonstrating the decay of information

In this section we demonstrate the decay of information using the Logistic Map as a model of the Quartic Map. In section 5.2.1 we define the components of the experiment: namely the climatology and the ensembles. Then we determine fitting parameters of the kernel dressing to build forecast densities. Kernel dressing was introduced earlier in chapter 3, section 3.6. Next in section 5.2.3, we compute measures of predictability such as the Empirical Ignorance, the Model Implied Ignorance and the Information Deficit (as defined in sections 3.3 and 3.4). Two methods of computing these measures (mD and mE) are presented and discussed.

5.2.1 Ensembles and Climatology

Details of experiment 5.D are given in Table 5.5. This yields the following datasets:

1. **Climatology** is built on $n_{icclim} = 2^{11}$ Initial Conditions drawn from U(0,1) as formed in the experimental design 5.A, which yields 2^{22} points evolved on system. The probability density for each bin of size of truncation between zero and one is calculated. For more details about how the climatology was calculated, see Appendix B.

2. 512 Initial Conditions form the in-sample set and another 512 the out-of-sample set. A total of n_{ic} IC were found using experimental design 5.D. This set is independent of that used as the climatology and constitutes **target forecasts (or outcomes)**.

3. **Ensemble members** are formed by taking the target forecast at ts_0 truncated to 2 decimal places, $+ r_m/100$, where r_m is U(0,1). There are m (32) members

in an ensemble and $n_{ens} = 1024$ ensembles (512 IC formed in-sample set and 512 formed out-of-sample set).

System	Quartic Map
Model	Logistic Map
r	4
μ	0.1
$TRAN$	128
n_{ic}	1024
n_{icclim}	2048
n_{ens}	1024
IC	U(0,1)
m	32
r_m	U(0,1)
ts (model)	15
ts (climatology)	2048
Truncation level	1/100
Size of climatology	2^{22}

Table 5.5: Experimental Design 5.D

Figure 5.4 illustrates an ensemble at lead time ts_0. The blue stars represent the initial condition ensemble and the red line the true initial condition (outcome).

The next step is to evolve all 1024 ensembles using the model and their 1024 target forecasts using the system forward 15 time steps.

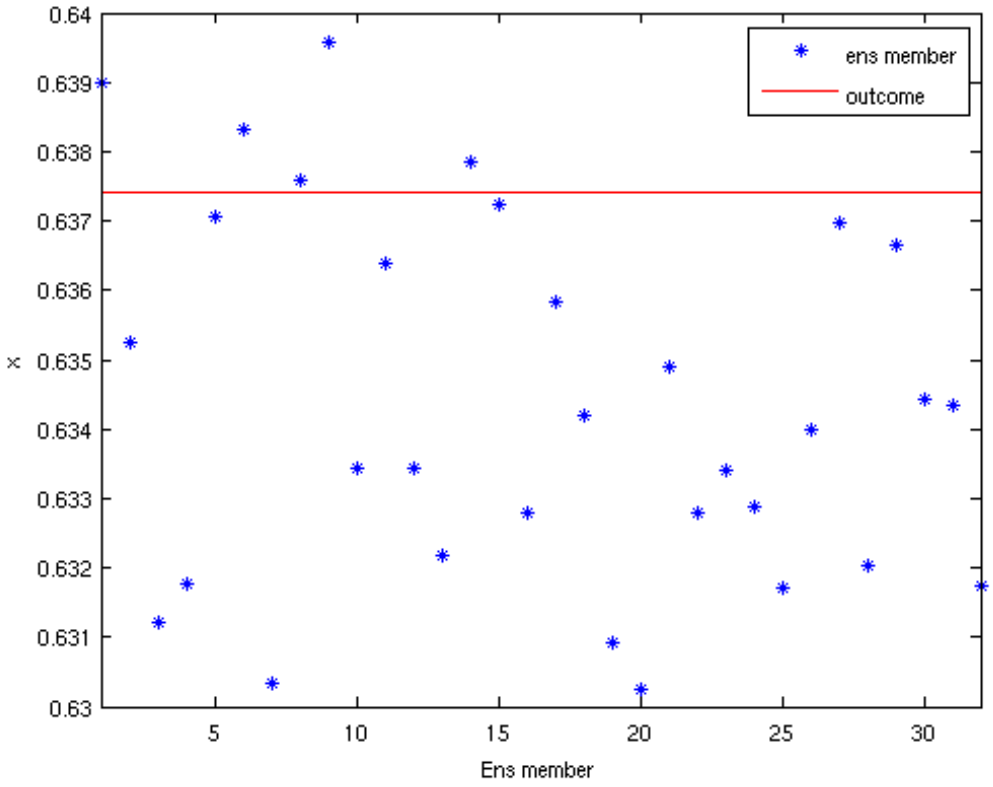


Figure 5.4: An example of initial condition ensemble for the experiment 5.D. The chart illustrates the location of 32 ensemble members. Each blue star represents an ensemble member, while the red line represents the outcome. All ensemble members will correspond to the same outcome.

5.2.2 Operational parameters for use in forecasting

We translate a given ensemble into a probability distribution using a combination of kernel dressing and blending with climatology as described in section 3.6. We estimate the operational parameters (kernel width σ and the weight α), which minimise Ignorance (the skill score). To do this we use the climatology and in-sample set described earlier in experiment 5.D.

Optimization of the kernel dressing and blending is done using Matlab optimization

routine fminsearch. To help ensure convergence to the global minimum, we use five different starting points from a range of 0 and 1.

Figure 5.5 shows the operational parameters (kernel width σ and the weight α) as a function of lead time[2]. The blending parameter α decreases from almost one at time step one to zero at time step 9, with the biggest drop between time steps 5 and 8. When α is large, the ensemble makes a major contribution to the forecasts. As for σ, it increases over the first 7 time steps, because the ensemble spreads out further over time. We note here the decrease of σ at time step 8. Later it fluctuates, but this does not matter as $\alpha = 0$ means that the forecast is equivalent to climatology.

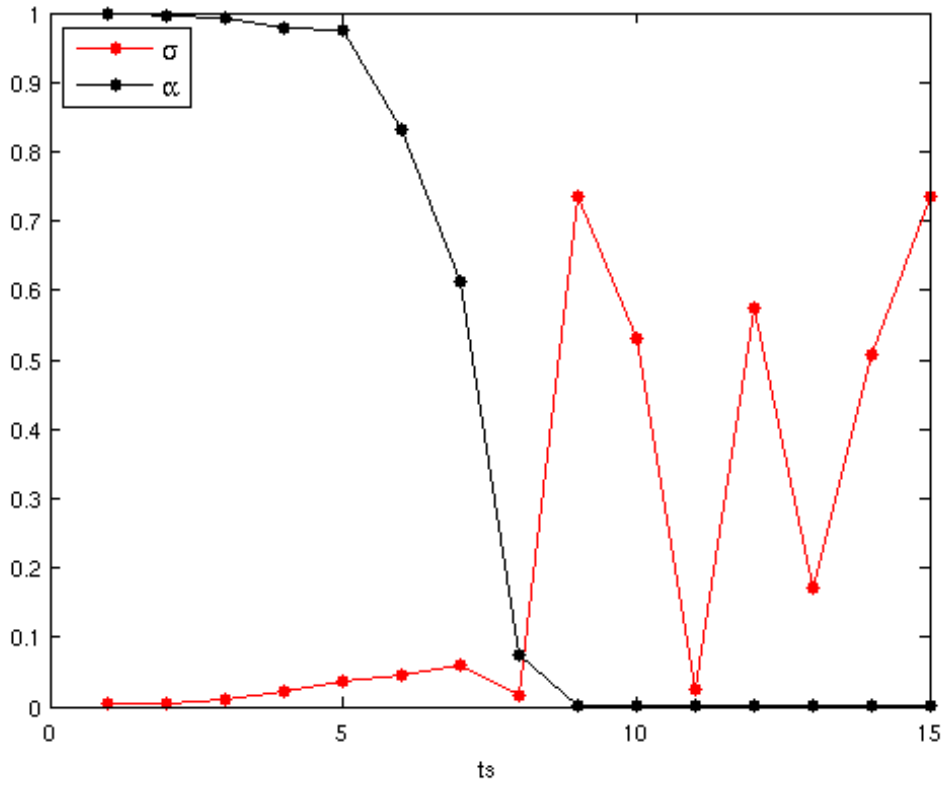


Figure 5.5: Operational kernel dressing parameters as a function of lead time in experiment 5.D[2]. The red line shows the kernel width σ and the black line shows the blending parameter α. For short lead times, α is close to one and hence the ensembles make a larger contribution to the final forecasts. Since the ensembles will spread out over time, a larger value of σ is required of larger lead times. Once α approaches 0, however, the kernel width becomes irrelevant.

5.2.3 Decay of Information

To demonstrate the decay of information we use data independent of those used to determine the operational parameters. We calculated the parameters over the forecast-outcome archive. To evaluate the forecast system we compute the Empirical

[2]An alternative way of looking at the right side of this plot is to consider it from the model selection perspective. In this case, from time step 9 the dressing and blending model is unnecessary and we go straight into climatology.

Ignorance, the Model Implied Ignorance and the Information Deficit over the out-of-sample set.

Selecting between two approaches to normalisation

Ignorance of the forecast

In this thesis, we express ignorance of a forecast relative to that of climatology. In the literature, there are two ways of computing this, depending on which approach is taken in considering the climatology, which was introduced in chapter 3.3.

The results of calculating the Climatological Ignorance are shown as a function of lead time in Fig.5.6. Values of methods D (mD) and E (mE) defined in equations 3.3 or 3.7, of section 3.3 respectively, are in the same range.

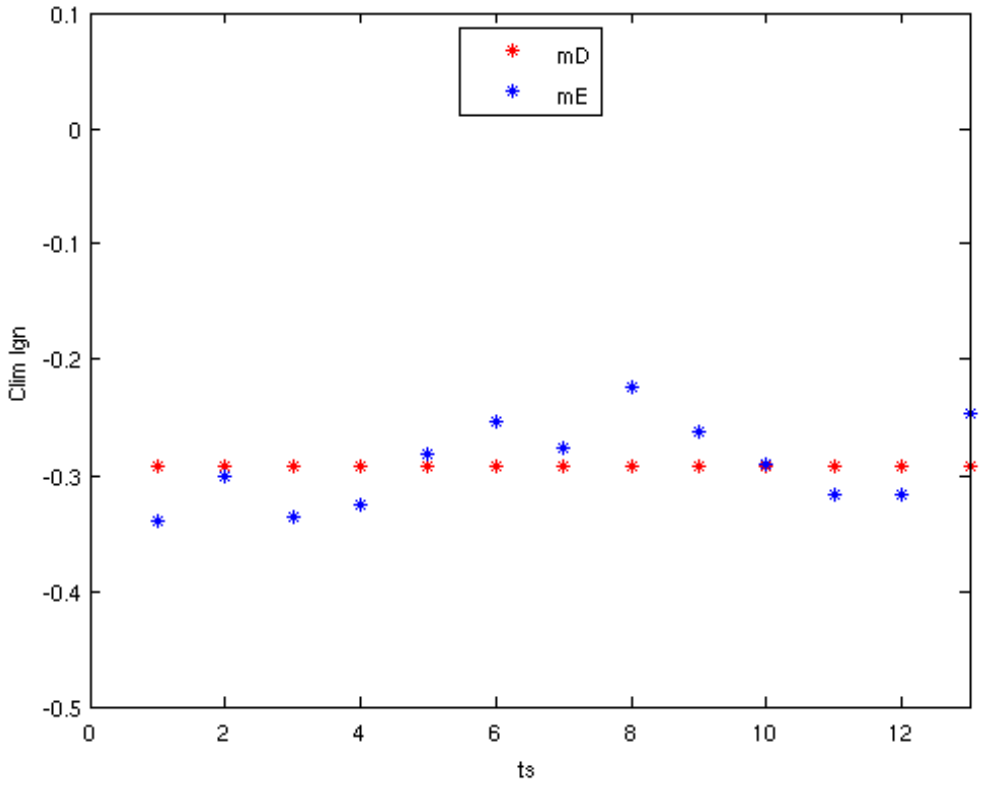


Figure 5.6: Climatological Ignorance using each method as a function of lead time. Method D (mD) assumes that target forecasts are random draws from the climatology. Method E (mE) is the Mean Ignorance when using climatology as a forecast. Values of the Climatological Ignorance calculated in methods D and E are in a similar range, but are not exactly the same.

There is not much difference in the Empirical Ignorance between both methods (see Fig.5.7), but note that mD overshoots the zero line. There is a huge contrast in the Ignorance of each forecast as shown in Fig.5.8. For method E (see right section on Fig.5.8), the Climatological Ignorance brings the ignorance of all forecasts down to zero from lead time 9. In method D, for longer lead times, the Ignorance stays in a similar range, but does not reach 0. Some forecasts have a positive ignorance score. This happens when the target forecast falls into part of the forecast density, which has low probability. This is expected to happen sometimes. This issue is illustrated

in Fig.5.9.

Mean Ignorance decays over time from around -5 bits at time step one to zero at time step 9 (see Fig.5.7). This means that for the first 8 time steps the forecast has skill, because it is different from the climatology. We notice the Mean Ignorance is positive for mD (see the left panel) at time step 8, though it is zero relative to the resampling interval.

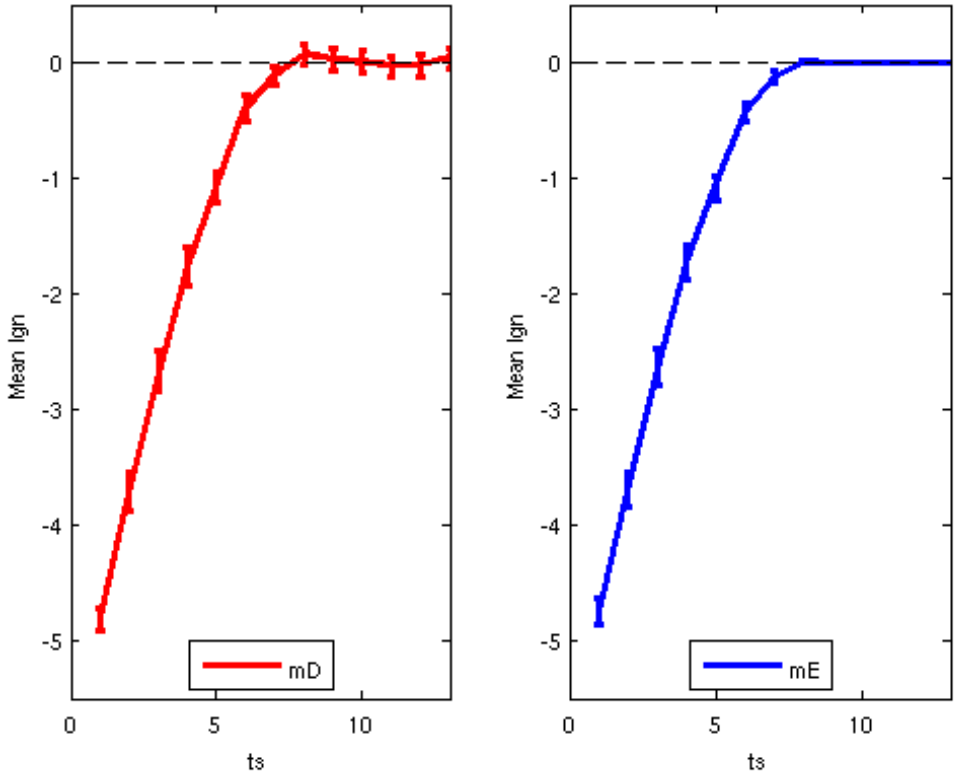


Figure 5.7: Mean Ignorance as a function of lead time for two methodologies (mD and mE), in experiment 5.D. It increases in the first 8 time steps and for method E at the 9th lead time converges to 0. When the Empirical Ignorance is zero, the forecast is equal to the climatology so it is not useful. There is no big difference between mD (left panel) and mE (right panel). Here we note that method D is positive at time step 8, though the resampling interval includes 0.

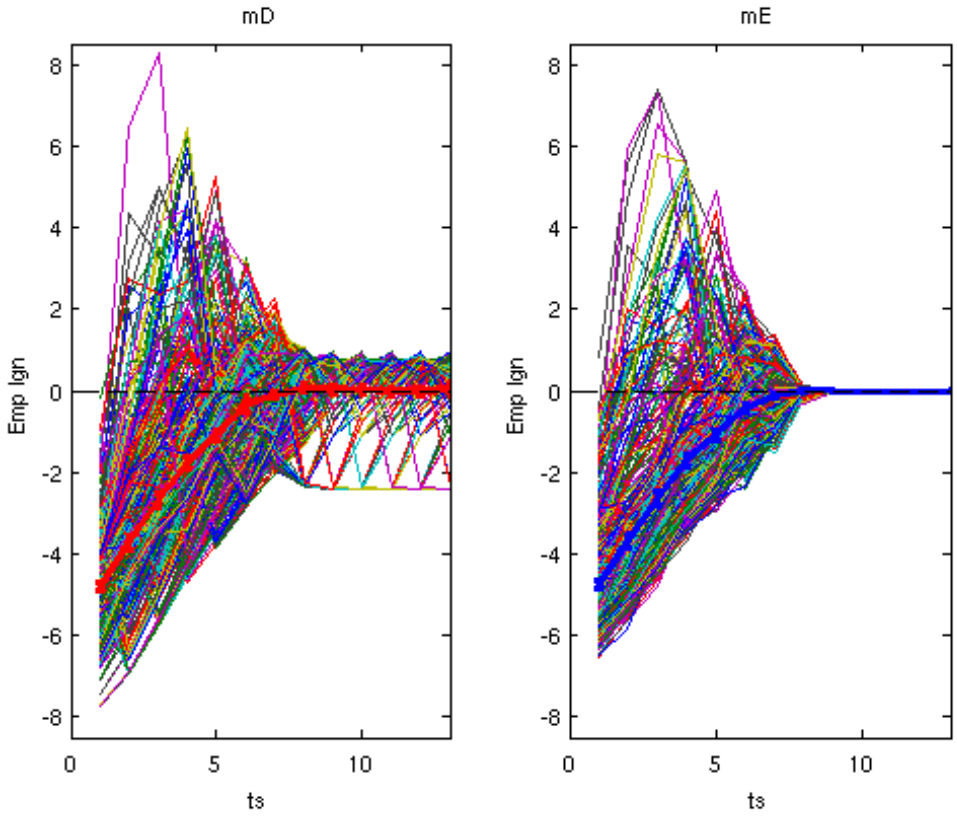


Figure 5.8: Empirical Ignorance calculated using two methodologies (mD and mE) for experiment 5.D. Each thin line shows the Ignorance of a different ensemble forecast, whilst the thick line shows the Empirical Ignorance. Climatological Ignorance in mE brings all forecasts to zero from time step 9 while in mD they do not converge. At initial time steps, the Ignorance is often much greater than zero in both methods. This is expected and explained in Fig.5.9.

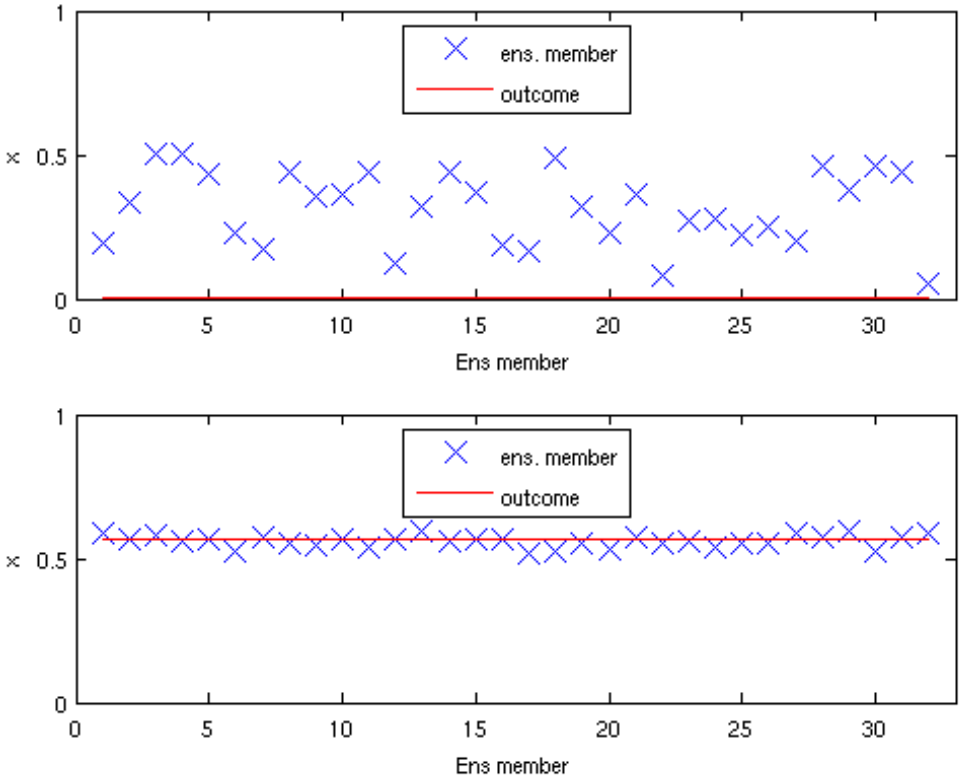


Figure 5.9: Examples of ensembles (blue crosses) and the outcomes (red lines) for a case in which the outcome falls outside of the range of the ensemble (top) and within the range of the ensemble (bottom). The ensemble at the top will tend to result in a positive ignorance score (see Fig.5.8).

Model Implied Ignorance

Model Implied Ignorance tells us what the expected ignorance would be if the model was perfect. The formula for calculating the Model Implied Ignorance was introduced in chapter 3.3. There are two methods of calculating these measures: using normalisation mD and mE. These give different results which are contrasted here.

Components that are subtracted when calculating the Model Implied Ignorance in mD and mE in experiment 5.D are shown in Fig.5.11. Method E converges to

method D from time step 9, that is when blending parameter α is 0.

There is some difference in the Model Implied Ignorance between both methods (see Fig.5.10); the climatological part that is subtracted (see equations 3.6 and 3.9) contributes to this difference (see Fig.5.11). Similarly to the Empirical Ignorance, this increases quantity over the first 8 time steps and then converges toward zero. We notice the Model Implied Ignorance is positive for mD at time step 8 (see the left panel of Fig.5.10), and the resampling interval does not include zero, suggesting it is significantly larger than zero.

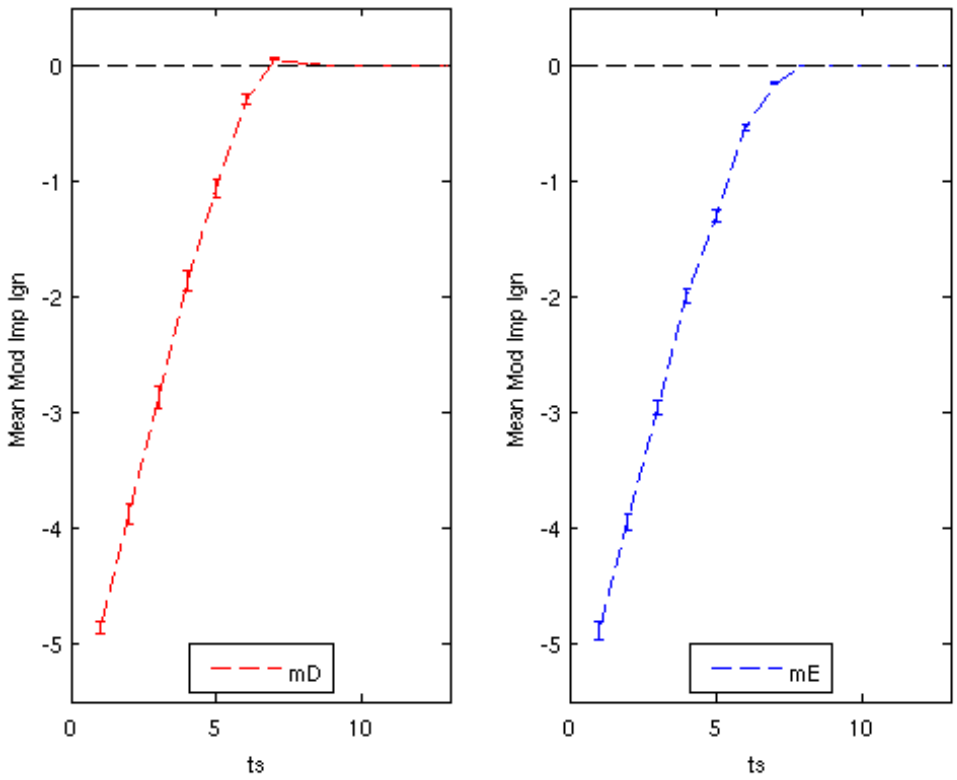


Figure 5.10: Model Implied Ignorance as a function of lead time for method mD (left) and mE (right) in experiment 5.D. Whilst method E always stays non-positive, this is not the case for method D.

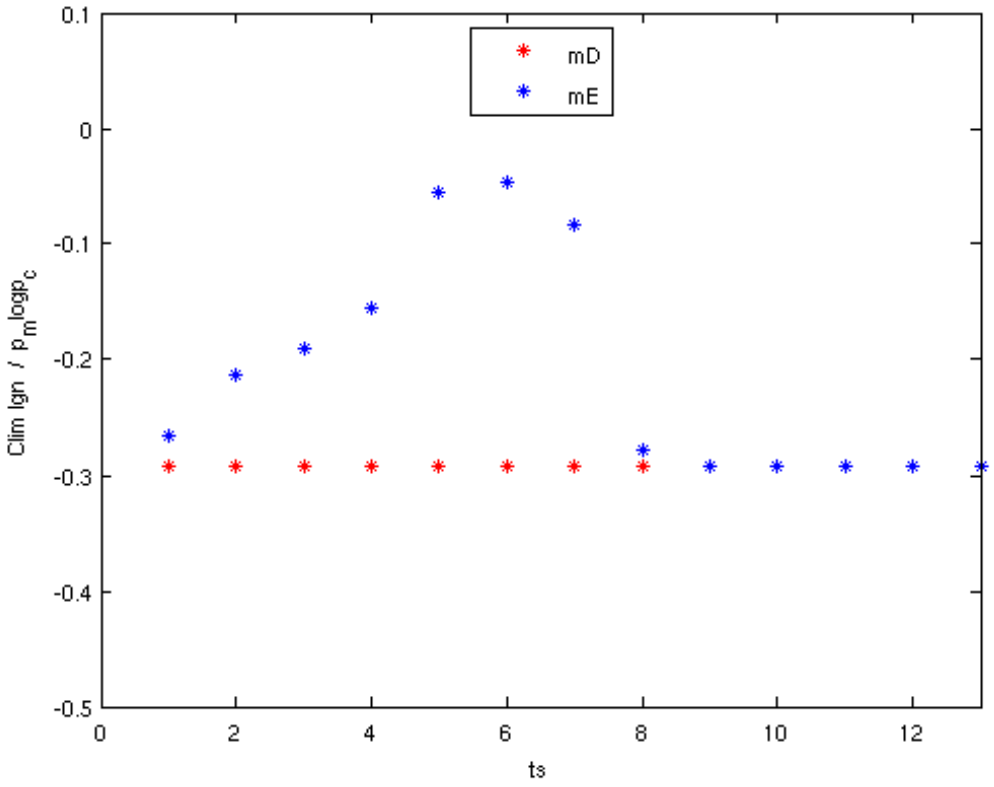


Figure 5.11: Components that are subtracted (see equations 3.6 and 3.9) when calculating the Model Implied Ignorance for method D (mD) and method E (mE) as a function of lead time. mE converges to mD after 8 time steps.

There is a contrast between methods mD and mE when calculating the Model Implied Ignorance. For example, between time steps 4 and 8 (see Fig.5.12) all forecasts in method E are nonpositive, but the Climatological Ignorance in mD takes some of the forecasts above zero. In both methods forecasts converge to zero from time step 9. This is because the blending parameter α is zero and the forecast is equal to climatology.

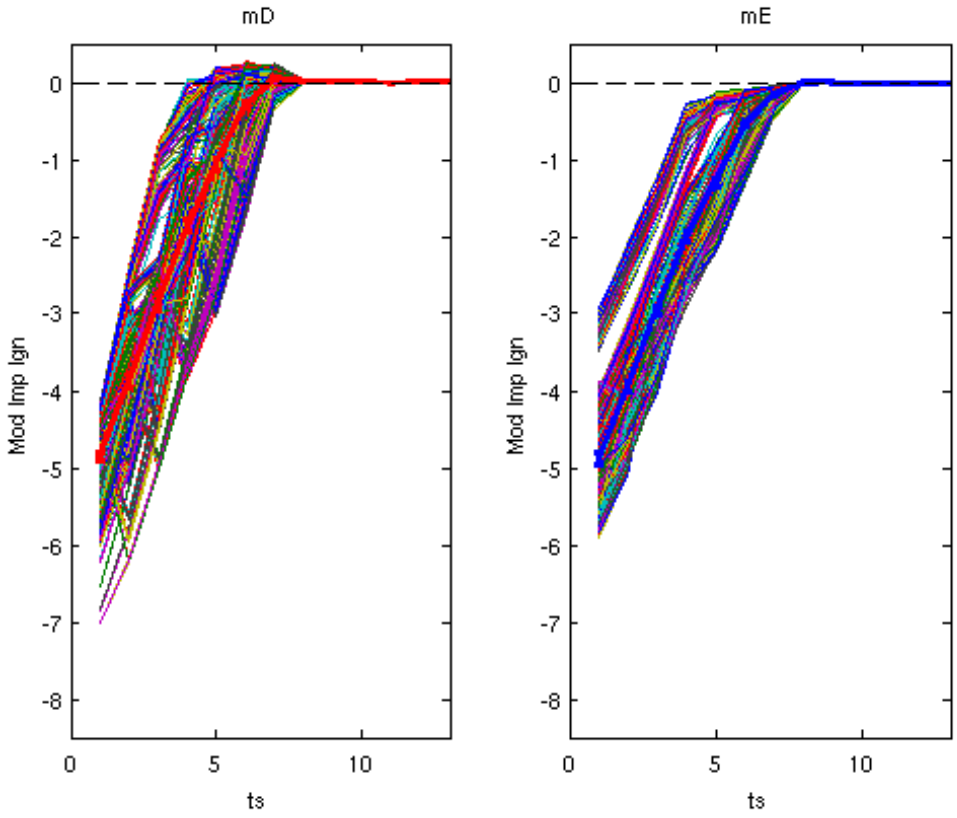


Figure 5.12: Model Implied Ignorance calculated using two methodologies (mD and mE), for experiment 5.D. Each thin line shows the Model Implied Ignorance for a different ensemble, whilst the thick lines show the mean over all forecasts at each lead time. In both methods, forecasts converge to zero from time step 9, since $\alpha = 0$. In mD the Climatological Ignorance takes the Model Implied Ignorance above zero in some cases.

Information Deficit

Fig.5.13 shows the Empirical Ignorance, the Model Implied Ignorance and the Information Deficit calculated using both method D and method E. At short lead times, both the Empirical Ignorance and the Model Implied Ignorance are low; around 5 bits. This means that our forecasts are, on average, 5 bits better than the climatology, that it they place 32 times more probability on the outcome than the climatology does. In both cases, the Ignorance is negative for the first 8 lead times

then converges to 0. The Information Deficit is positive when the Model Implied Ignorance is better than the Empirical Ignorance. In that case the dashed line in Fig.5.13 would be below the solid line. The Model Implied Ignorance is expected to be less than the Empirical Ignorance, making the Information Deficit positive. When the Empirical Ignorance and the Model Implied Ignorance are the same, the Information Deficit is zero.

Here, in normalisation mE (see Fig.5.14), the Information Deficit is positive for short lead times and eventually drops to zero, when the Model Implied Ignorance and the Empirical Ignorance both converge to zero. Using normalisation mD (left panel) the Information Deficit is positive for the first 4 lead times, but then drops below zero. These negative values are surprising since the Information Deficit is expected to be positive. The reason for this occurrence is explained in the next section, 5.2.4, in which shortcomings of method D are exposed.

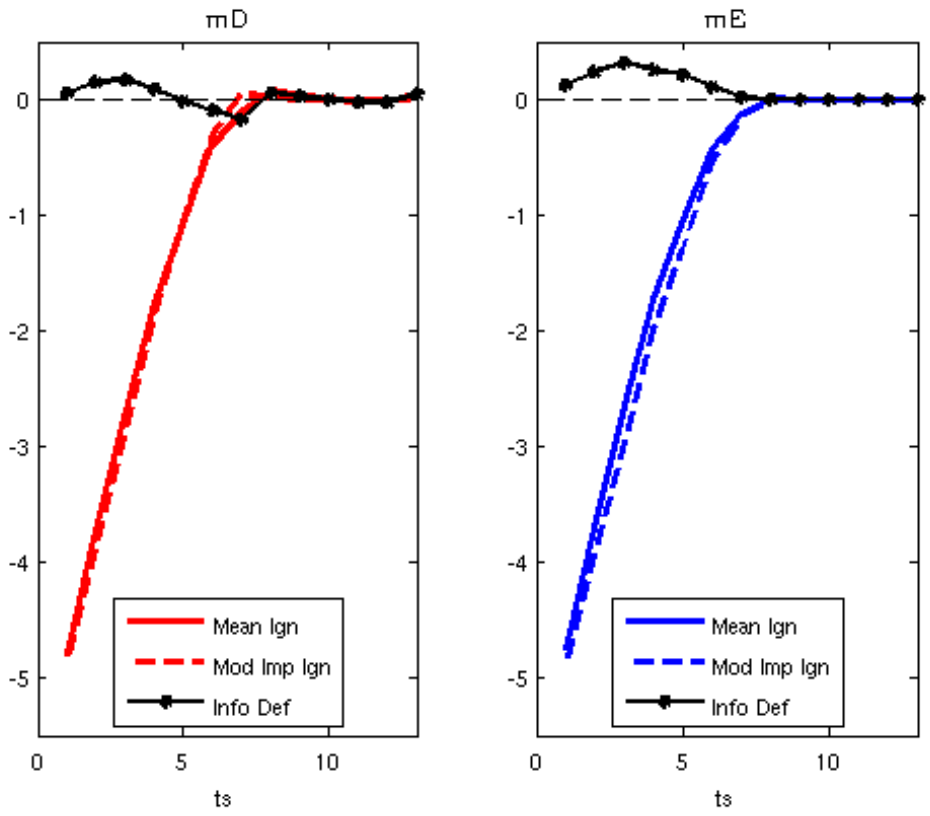


Figure 5.13: Empirical Ignorance (solid line), the Model Implied Ignorance (dashed line) and the Information Deficit (black line) as a function of lead time for methods D (left) and E (right). Whilst the Information Deficit is non-negative for method E, this is not the case for method D.

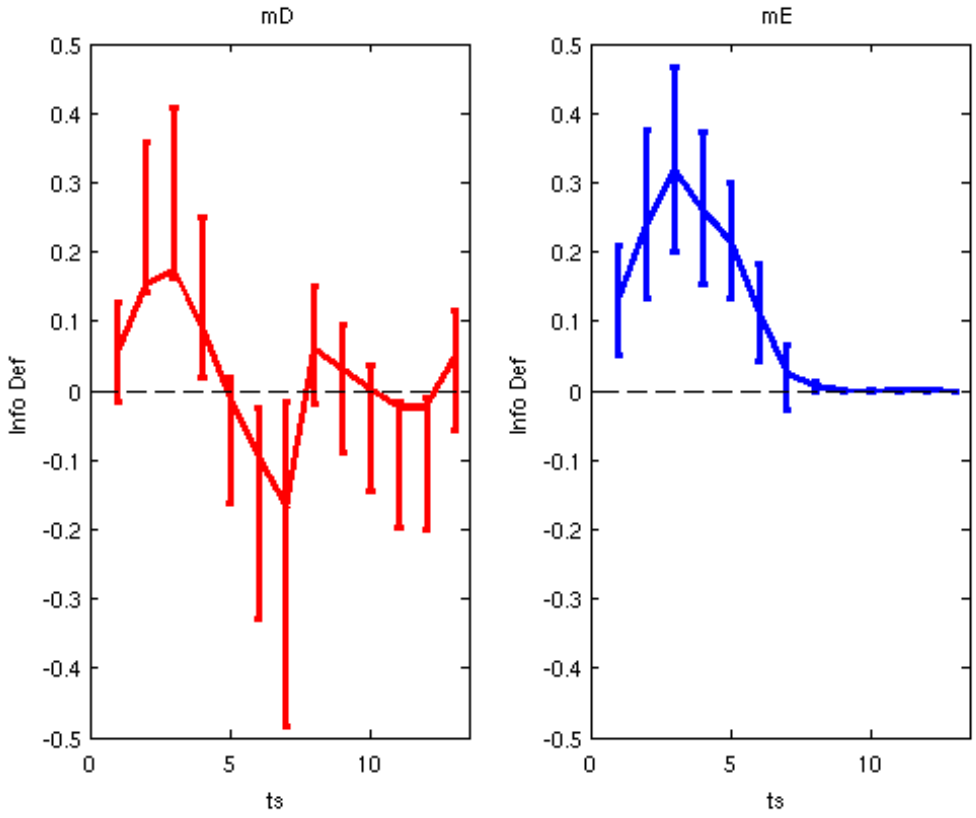


Figure 5.14: Information Deficit calculated using method D (left) and E (right) with 95% resampling intervals as a function of lead time. For method D, there is significant evidence that the Information Deficit is negative in 4 lead times.

5.2.4 Shortcomings of normalisation method D

The Empirical Ignorance, and the Model Implied Ignorance, and hence the Information Deficit, can be calculated using normalisation by method D or method E. Method E uses the ignorance when the climatology is used as a forecast for the actual outcomes whilst method D uses the expected ignorance of an outcome drawn randomly from the climatology. We now summarise a number of shortcomings of method D.

1. Consider a forecast formed using blending in which the blending parameter α is equal to zero. Since the forecast is equivalent to the climatology, it might be expected that the ignorance will be zero in all cases. Defining S_{clim} to be the expected climatological ignorance is confusing and may even give the misleading impression that forecast skill is present for some forecasts. This may lead the forecaster to attempt to separate forecasts with and without skill, when there is, by construction, zero skill.

2. Method mD first assumes that outcomes resemble iid random draws from the climatological distribution when, in reality, serial correlation in the outcomes may mean that this is not the case. [55]

3. In general, the climatological distribution must be estimated from past observations. This means that, strictly speaking, the climatological distribution estimated by the forecaster is not the precise distribution from which the outcomes were drawn. Method mD acts as though the estimated climatological distribution is perfect; mE does not.

For the reasons stated, the use of method D is not recommended and hence, for the rest of this thesis we use method E.[3]

5.3 Logistic Map as a model for alternative systems

Next we repeat the experiment 5.D from the previous section for different values of μ in the system. We consider four different values of μ. When $\mu = 0$, the model is

[3]The originators of these different normalisation methods agree with the conclusions of this section (H. Du, L.A. Smith and E. Wheatcroft, 2016 personal communications).

perfect[4]. When $\mu = 0.01$ we have shown that the model is close to perfect. When $\mu = 0.1$ or $\mu = 0.5$ there is a larger difference between the model and the system. We will refer to these experiments as Experiment 5.E. The details are listed in Table 5.6.

System	Quartic Map
Model	Logistic Map
r	4
μ_1	0
μ_2	0.01
μ_3	0.1
μ_4	0.5
$TRAN$	128
n_{ic}	1024
n_{icclim}	2048
n_{ens}	1024
IC	U(0,1)
m	32
ts (model)	13
ts (climatology)	2048
Truncation level	1/100
Size of climatology	2^{22}

Table 5.6: Experimental Design 5.E

Here we are investigating the value of the ignorance and the Information Deficit in terms of demonstrating the decay of predictability. The Information Deficit has been defined in a previous section of the thesis (3.4). If the Information Deficit is negative, the model is under confident. This suggests we are missing something obvious or have made a silly mistake and the Information Deficit might prove a useful tool to indicate such problem. If the Information Deficit is positive, the model is overconfident.

Initially, after experiment 5.E had been conducted, we found that the Information

[4]Note that when μ=0 the system is equivalent to the Logistic Map with r=4, which has a known natural measure. This is discussed in Appendix B.

Deficit was negative. This was surprising as the Model Implied Ignorance is expected to be better than the Empirical Ignorance. We reviewed the method of calculating the forecast and found an error when calculating the Empirical Ignorance, the Model Implied Ignorance and the Climatological Ignorance. A feature of standard kernel dressing is that it does not account for boundaries. System states of the Quartic Map are known to be bounded above by one, and below, by zero. The integral of the forecast density over the range (0,1) should therefore be one. This was not always the case when the kernel smoothing function was applied. The forecast was putting the probability outside the range of the map, so some of it was assigned to areas in which outcomes could not fall. This was fixed by dividing each forecast by a normalisation constant, but the Information Deficit was still negative for time steps 5, 6 or 7 in each considered case of the system. We checked, that for these time steps there was a decrease in the blending parameter σ (bandwidth of the kernel). Further examination showed that operational parameters were not optimised with respect to the normalised Ignorance. As a result of all of these corrections, all of the results are changed; the Model Implied Ignorance is now better than the Empirical Ignorance and the Information Deficit is never negative (see Fig.5.15). This investigation demonstrated that the Information Deficit can be a useful tool to capture both errors in the code or costly assumptions in the forecast methodology.

Figure 5.15 shows the corrected Information Deficit for the first 13 time steps in experiment 5.E. Each colour indicates a different parameter μ of the system. When the model is perfect (μ=0) or the model imperfection is low (μ=0.01) the Information Deficit is in a range of around 0.1 to 0.3 for time steps 2-6. The Information Deficit tends to be larger when the imperfection is higher (when μ=0.5) for the first 4 time steps. The black dotted line shows the Information Deficit for an unblended model

[5], when α is large (0.9922) and fixed for all time steps. Other models have a lower Information Deficit after time step 4, which is due to blending with climatology. The Information Deficit converges to zero at time step 8 when μ=0.5, in time step 9 for μ=0.1, in time step 11 when μ=0.01 and in time step 12 for μ=0. The blending parameter $\alpha = 0$ at these time steps and then the forecast and climatology coincide.

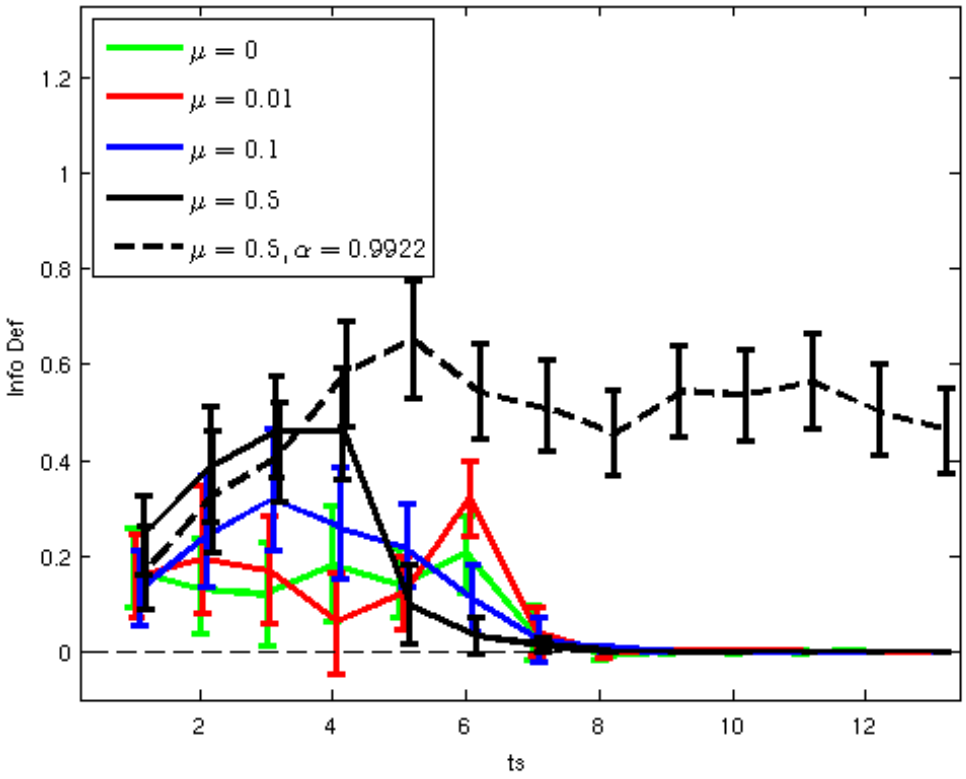


Figure 5.15: The Information Deficit as a function of lead time for different values of μ. The Information Deficit is positive and larger, in general, for the highest level of model imperfection. The larger the μ, the quicker the Information Deficit reaches zero. This is because the blending parameter α decays to zero quicker. The dashed line shows the Information Deficit when $\mu = 0.5$ and the model is not blended with climatology. The Information Deficit tends to be the largest in this case and never reaches 0.

[5]Technically more accurate terminology would be 'a not fully blended model'. A model is unblended when $\alpha = 1$.

Next, we look more closely at the two components that go into the calculation of the Information Deficit: 1) the Empirical Ignorance and 2) the Model Implied Ignorance. The first tells us how much skill the model has and the second gives the expected Ignorance, were the outcome drawn from the forecast itself. Fig.5.16 shows these values for the case of $\mu = 0$, $\mu = 0.01$, $\mu = 0.1$ and $\mu = 0.5$.

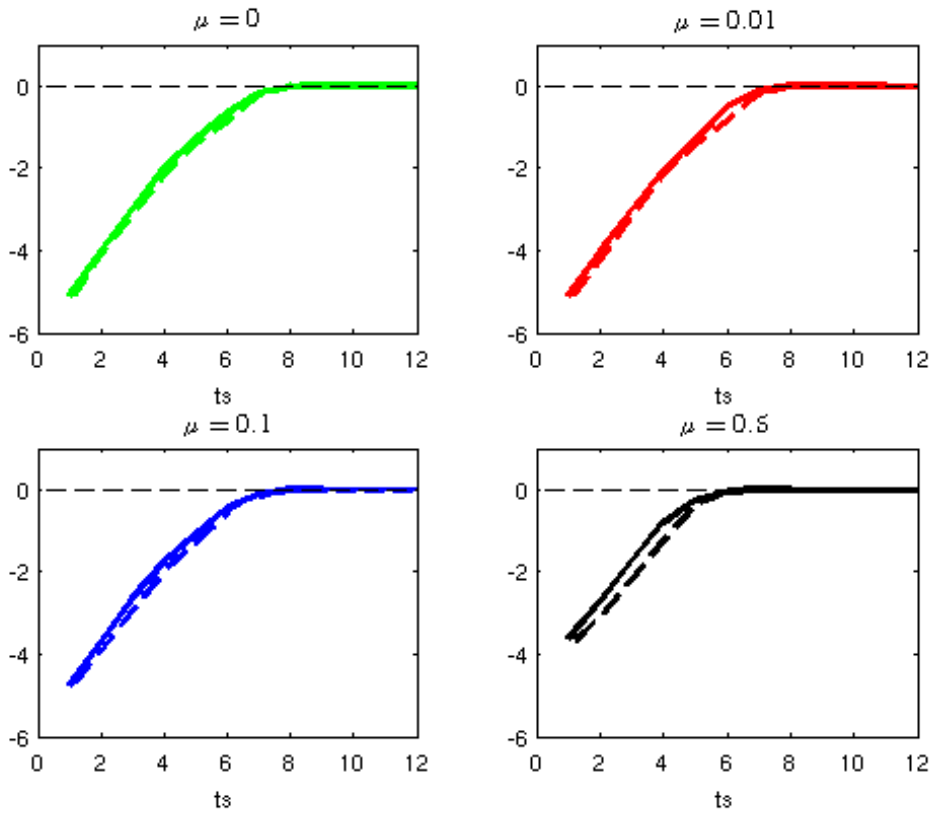


Figure 5.16: Empirical Ignorance (solid line) and the Model Implied Ignorance (dashed line) as a function of lead time for different versions of model. The Model Implied Ignorance is better than the Empirical Ignorance. The gap between the two types of Ignorance is the largest for the more imperfect model (when μ is 0.5, black lines). The smaller the μ, the longer it takes for the Empirical Ignorance and the Model Implied Ignorance to converge to 0. Both measures of predictability increase as μ decreases.

At first, the Empirical Ignorance (solid line) increases very quickly. The decay of

information then slows down as it approaches zero. At longer lead times, the forecast distribution approaches climatology and no longer contains useful information. The smaller the μ, the better the Ignorance relative to climatology and the longer before the Ignorance decays to zero.

The Empirical Ignorance in Fig.5.16 was computed on the out-of-sample set. Operational parameters for each value of μ were calculated on the in-sample set. If we plotted the Empirical Ignorance estimated on the in-sample and out-of-sample sets on the same chart both curves would be hard to distinguish. The Empirical Ignorance is occasionally better for μ=0.01 than for the Perfect Model. Since it is the system that changes there is no contradiction in the μ=0.01 being more predictable than the μ=0 case.

Fig.5.17 shows Ignorance for each individual forecast for μ=0.5 and μ=0.01. The Climatological Ignorance brings the Ignorance of all forecasts to zero from time step 8 when $\mu = 0.5$ and 9 when $\mu = 0.01$, because α=0 at these lead times and the forecasts and climatology coincide. Some forecasts have positive ignorance. This usually appears to happen when out-of-sample ensemble members are far from the target forecast (Fig.5.9) and the kernel width is 'too narrow' to be effective in these specific spaces. In that case, the Empirical Ignorance is sometimes a good deal greater than 0. For $\mu = 0.01$ (see left half of Fig.5.17) there are some forecasts with the Empirical Ignorance greater than 10 at time step 3. The range of values of the Empirical Ignorance is smaller for μ=0.5 than for μ=0.01.

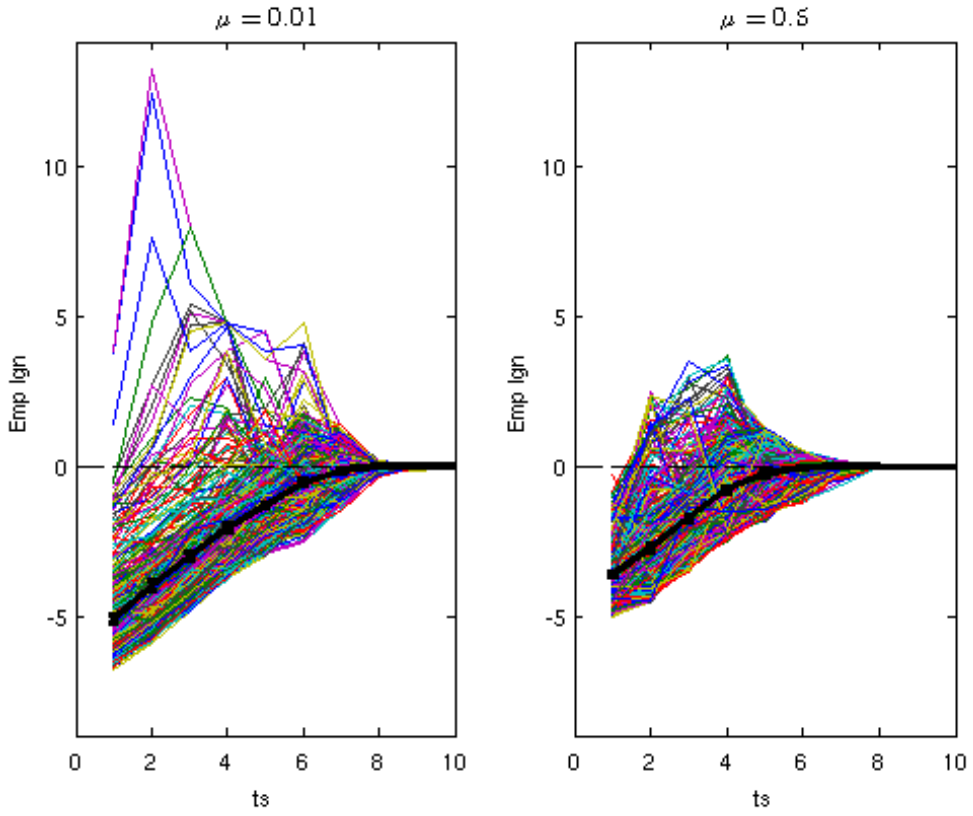


Figure 5.17: Empirical Ignorance when μ=0.01 and 0.5. Each thin line represents a different forecast. The black thick curve shows the Mean Ignorance over all forecasts with the error bars representing 95% resampling intervals of the mean. For longer lead times, the Ignorance converges to 0. There are some bad forecasts (those with the Empirical Ignorance well above 0) at the beginning of the simulation for each of these two cases of μ. The score forecast is much worse for μ=0.01 than 0.5, with the Empirical Ignorance being greater than 10 (see chart on the left). This happens when, for example, ensemble members are far from the target forecast and the kernel width is too small to capture them (see Fig.5.9). If operational parameters used to calculate the forecast are modified as discussed in the text, the number of 'bad' forecasts decreases (see Fig.5.18).

If we calculate the Empirical Ignorance of the in-sample set in case of μ=0.01, the number of 'very bad' forecasts at time step 3 is reduced from 6 to 3. Here we define 'very bad' forecasts to be those with the Empirical Ignorance greater than 5 bits; this threshold is, of course, arbitrary. Increasing the kernel width by a factor of

two or decreasing α by one eighth at each time step reduced forecasts with bad predictability significantly in both cases (see Fig.5.18). For example, here at time step 2, there are no very bad forecasts when σ is twice larger or when α is decreased.

The whole point of using an ensemble is to compensate for the odd, bad forecast. It is understood that individual forecasts should never be used separately in any case. Here we change the blending parameters α and σ in an ad-hoc way, just to explore what the impact of that change would be.

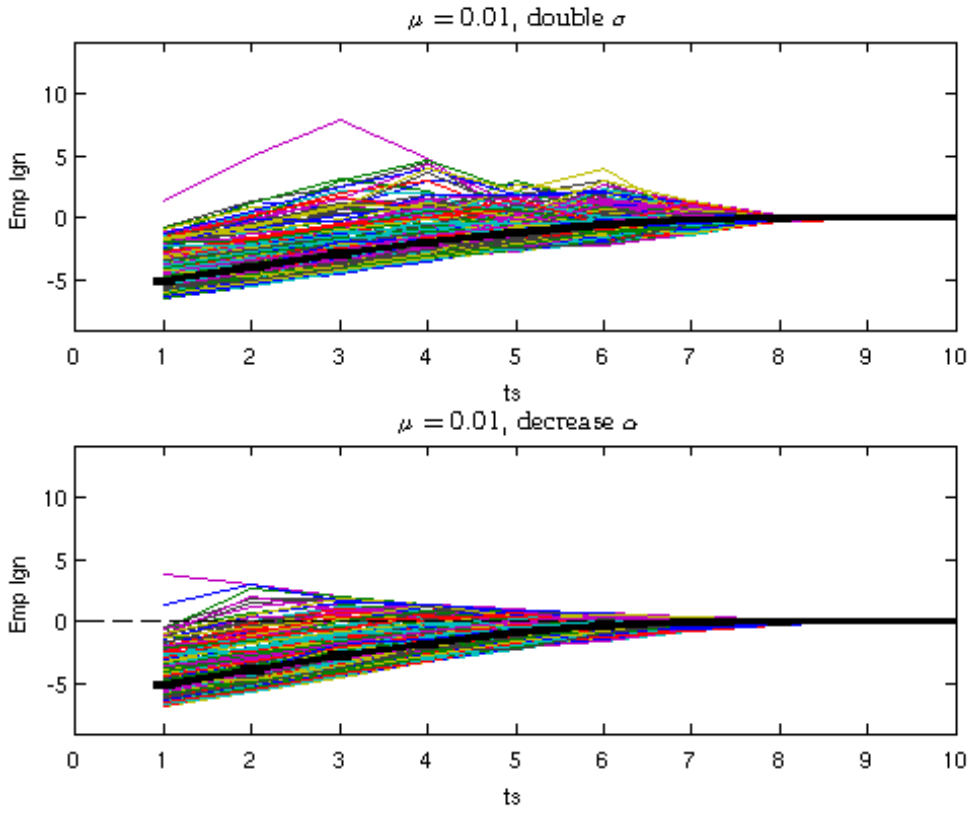


Figure 5.18: Empirical Ignorance by time step for μ=0.01 calculated with modified α and σ. This should be compared with the left panel of Fig.5.17. Doubling of the kernel width σ or decreasing α by one eighth for each time step reduced the number of 'bad' ensemble forecasts.

5.4 Conclusions

Properties of the model (Logistic Map) and the system (Quartic Map) have been compared for different levels of imperfection in the model caused by tweaking the parameter μ which lies on the interval $(0, 1)$ and is equal to the system when $\mu = 0$. The similarity between the invariant measures of the model and the system was demonstrated in Fig.5.1 and explored in more detail by comparing the relative entropy of the two distributions (Fig.5.2). From this, it was concluded that the system behaves notably differently from the model for values of μ greater than approximately 0.06. This conclusion is reinforced by the dynamical similarity of the behaviour of the model and the system shown by examination of the bifurcation diagrams presented in Fig.5.3. To our knowledge, this is the first exploration of this kind for the Quartic Map.

Next, we demonstrated the decay of information using the ignorance score to evaluate the amount of useful information in the model relative to a baseline climatological forecast. We showed, using the Quartic and Logistic Maps how the blending parameter α eventually approaches zero and thus the information contributed by the model also approaches zero until no weight is placed on the model at all.

We then considered the use of Ignorance scores and compared two alternative ways of calculating the Empirical Ignorance and the Model Implied Ignorance using normalisation method D and method E. We demonstrated the calculation of the Information Deficit with reference to the Logistic Map as a model of the Quartic Map. Important shortcomings of method D, novel to this thesis, were identified and it was concluded that method E should be used exclusively.

Next, we performed a novel exploration of the use of the Information Deficit and

how the forecast never performs as well as it would, were it truly the distribution from which the outcome was drawn. We also discussed that, when the blending parameter is zero (that is, when no weight is put on the model and the forecast is simply the climatology), by construction, the Information Deficit is always precisely zero.

Finally, we described new attributes of the Information Deficit and demonstrated its use as a diagnostic tool in identifying areas of improvement in the forecasting methodology. The Information Deficit helped to capture a number of different shortcomings: (i) Forecast density was routinely being placed outside of the possible range of the outcome, (ii) bugs in the code, (iii) errors arising from the default settings of some Matlab functions (for example, using too few points in the numerical integration scheme).

Chapter 6

Modelling the real world: Predictability in probabilistic modelling of El Niño

In this chapter we apply measures of predictability developed earlier in this thesis to a real-world model, which is much more complex than the mathematical low-dimensional models discussed in the previous chapter. Here, our research concerns the Cane-Zebiak model (C-Z model), which is used to model El Niño events [1]. El Niño has widespread socio-economic impacts hence improved predictions of its behaviour would be of great value to society in terms of reducing damage costs [42, 41]. This has motivated us to study predictability in the context of the C-Z model. I received a scholarship [1] from the London School of Economics and Political Science (LSE) to learn about this model at Columbia University in New York, where the C-Z model was created and has been developed.

[1] Sep 2014-Nov 2014 Funded Partnership PhD Mobility Bursaries for LSE PhD students.

In section 6.1, we define El Niño and explain its climate and socio-economic effects both in the Pacific region and across the world. In section 6.2, we focus on the C-Z model and demonstrate that it is an appropriate model for the application of the non-linear methods which we have been developing in this PhD thesis. Next, an original experiment in which the decay of information is demonstrated is carried out. The predictand of our interest in this chapter is the Niño 3.4 index. Results on the predictability of the model in a Perfect Model Scenario are discussed in section 6.3, and in an Imperfect Model Scenario in section 6.4. Here, as in the previous chapter, we describe how the Information Deficit unearth practical difficulties in section 6.3.4. Defining a climatology in the C-Z model is shown to present different challenges than the mathematical model case (more about this is presented in section 6.3.2). Issues arising from not having a representative climatology are explained in section 6.5. Section 6.6 summarises the conclusions and identifies which results are new.

6.1 El Niño

6.1.1 What is El Niño?

Roughly speaking, El Niño is defined as the warming in the sea surface temperature (SST) when compared with the averaged value, from the coasts of Peru and Ecuador to the equatorial central Pacific Ocean [76]. The term was introduced by Peruvian fishermen, who recognised that the appearance of large masses of warm water had impacted their businesses, resulting in low catches. Since the warming begins around Christmas, they called it El Niño, which is Spanish for the 'Christ Child'.

There are four different Niño Regions in the Pacific (see Fig.6.1). This research

mostly concerns the Niño 3.4 region bounded by 5°N to 5°S, and 170°E to 120°W. The Niño 3.4 index may be regarded as the most appropriate ENSO index to use [12], because of the dominant effect of SST variability within this region on the global climate.

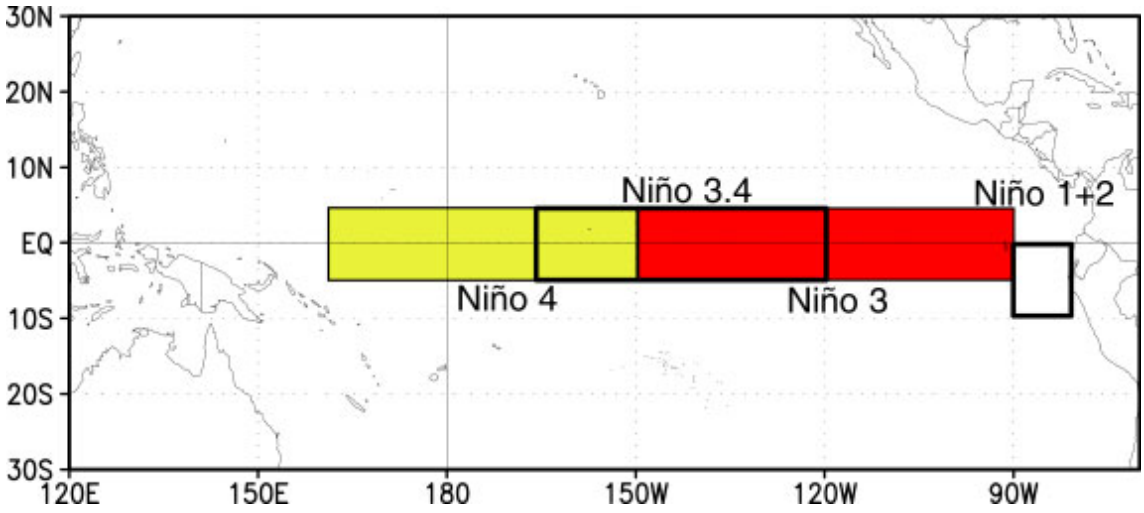


Figure 6.1: Graphical illustration of the four Niño regions in the tropical Pacific, by NOAA Climate Prediction Center [7]. These are the regions used by the C-Z model.

Different countries use different indicators to identify the phenomenon in the tropical Pacific. For example, the National Oceanic and Atmospheric Administration (NOAA) uses the Ocean Niño Index (ONI), which defines an El Niño event as one in which the 3-month average sea surface temperature (SST) anomaly in the Niño 3.4 region is equal to or exceeds 0.5°C for 5 consecutive periods in a row. Further, based on the magnitude of the SST change, weak (with a 0.5-0.9°C SST anomaly), moderate (1-1.4°C), strong (1.5-1.9°C) and very strong (above 2°C) events can be derived [4]. Fig.6.2 shows the ONI since 1950. Warm phases based on the threshold 0.5°C are above the first red line. The most recent El Niño conditions occurred at the beginning of 2015. Warm phases in 1982-3, 1997-8, and 2015-16 belong to the very strong category.

El Niño is the warm phase, while 'La Niña' is the cold phase of a larger irregular and recurrent climate pattern called the El Niño-Southern Oscillation or ENSO. El Niño occurs on average every 4 years, though it can occur between 2 and 7 years apart. Larger events generally start around summer, reach their peak near the end of the year and then weaken before the following summer during May-July. The length of the phase is therefore around a year [90]. The El Niño evoluation pattern described above is depicted in Fig.6.2 (Courtesy of NOAA [7]).

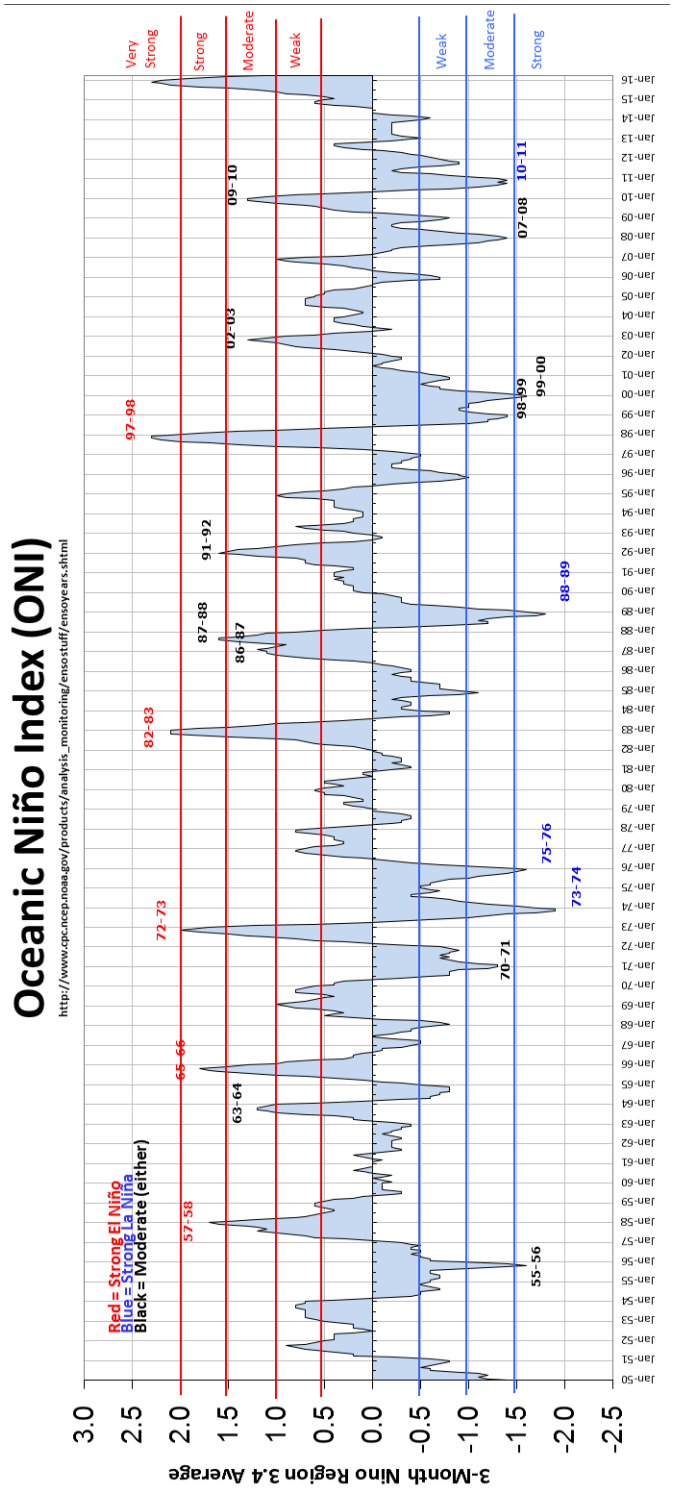


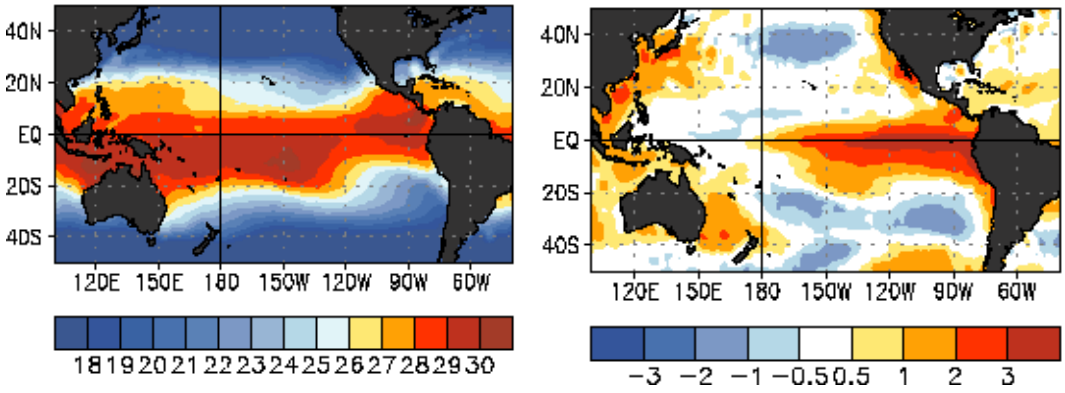
Figure 6.2: 3-month average sea surface temperature (SST) anomalies in the Niño 3.4 region. The figure illustrates irregularity in the length and occurrence of the events since 1950. The latest warm phase in 2015-16 was very strong with the anomaly above the 2°C threshold. (Courtesy of NOAA [7]).

El Niño affects not only the climate of the Pacific Ocean region, but is also associated with different temperature and precipitation patterns around the world, so called 'teleconnections' [43]. These can have various economic consequences. We will go into greater detail about this later in the chapter.

6.1.2 El Niño in the Pacific

El Niño appears in the tropical Pacific. Fig.6.3 illustrates Ocean Temperature Patterns, while Fig.6.4 shows precipitation in the first quarter of 1998 during which a very strong El Niño occurred over that region. Across the eastern half of the tropical Pacific, exceptionally warm waters (3-4°C warmer than normal) coincided with increased rainfall and thunderstorm activity. The right panel of Fig.6.4 shows regions with twice the average tropical rainfall (dark green area). In contrast, over the western equatorial Pacific rainfall was up to 600-800 mm less than normal (dark grey/ yellow on the left chart), which suggests a large deviation from the average conditions of that region.

This extreme dryness led to a series of major uncontrolled wildfires across Indonesia, Malaysia, and Northern Australia [42]. El Niño completely changes this region's climate.



(a) Total (°C) (b) Departures (°C)

Figure 6.3: Fig. a) shows average ocean temperature in the Pacific between Jan-Mar of 1998 during a very strong El Niño. Fig. b) relates to that period and shows departures[2] from the long-term average. We see there, that temperatures were above average by 3-4°C across the eastern tropical Pacific. (Courtesy of NOAA [7]).

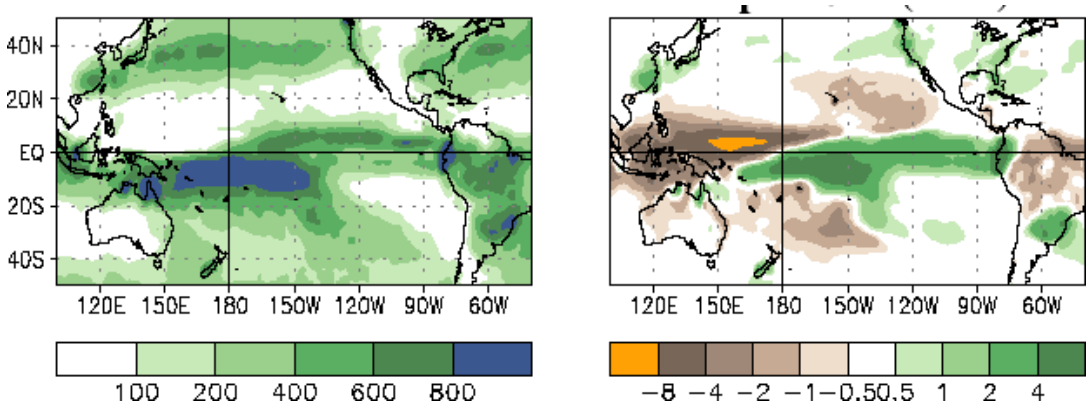


Figure 6.4: Map in the left panel illustrates total rainfall in Jan-Mar 1998, during a very strong El Niño. The heaviest rainfall at that time is shown by the darker green and blue colours. Analogically to Fig.6.3 the map on the right shows departures (x100)mm from the average. The dark green in the right picture indicates 400 mm more than average, so the rainfall was double the average in this region. By contrast, the dark grey region of Indonesia has, on average, 800 mm and in Jan-Mar 1998 had no precipitation at all, resulting in extreme dryness. (Courtesy of NOAA [7]).

6.1.3 El Niño and the World

El Niño is thought to result in changes to precipitation and temperature patterns in many parts of the world [86, 87, 61, 48]. In some regions the risks of floods and cold weather is increased whilst, in others, there is an increase in the risk of wildfires, droughts and heat waves. Impacts associated with the phenomenon depend on the season. It must be stressed that while these effects are more likely to occur, they are not guaranteed. Rather than going into detail about the effects in each region, we simply summarise some of the most likely impacts of El Niño around the globe. The two pictures in Fig.6.5 show global temperature and precipitation patterns associated with El Niño during winter (Dec-Feb) and summer (Jun-Aug).

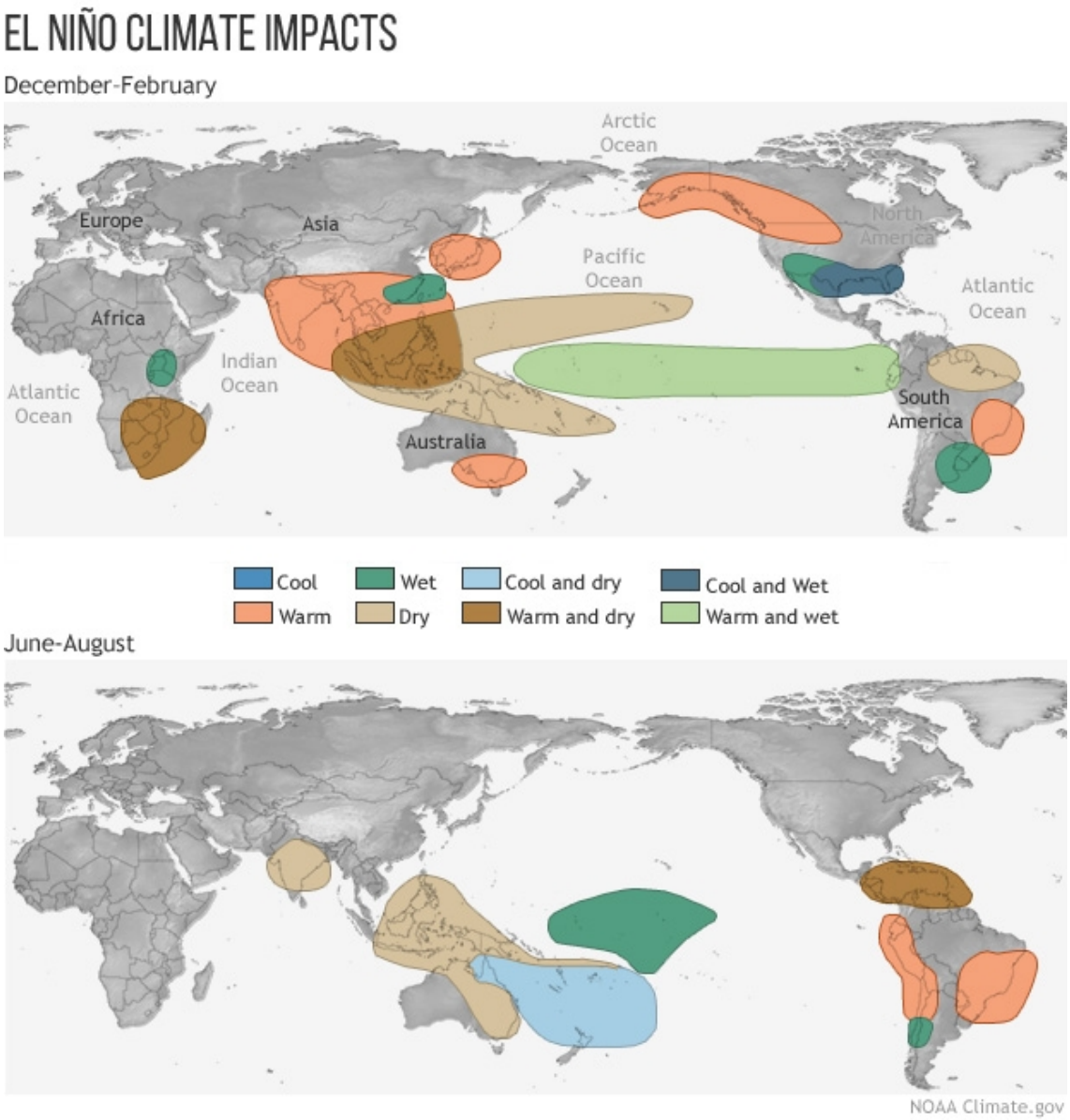


Figure 6.5: Effects of El Niño on global climate in winter (top) and summer (bottom) by NOAA [7]. Whilst these effects are more likely in El Niño years, they are not guaranteed.

During winter in an El Niño year, wet and warm conditions in coastal Ecuador and Peru are likely to occur. The phenomenon may also bring warmer winter temperatures across southern Alaska, western Canada and the northern US. Excess rainfall

is commonly observed with El Niño in eastern equatorial Africa and the southern US, while drier than normal conditions become more likely over northern South America, Central America and southern Africa between December and February. El Niño may also bring lower temperatures along the Gulf Coast of the US.

During summer, a common El Niño effect is that Indonesia and parts of northern South America tend to be drier than average, as does eastern Australia, Central America and India. Excess rainfall in central Chile becomes more common with El Niño between June and August. There is a tendency for above-average temperatures at that time of the year in southeastern and northestern South America and Central America, while the northern part of New Zeland may experience cooler and drier than average conditions.

The overview above shows that El Niño works both locally in the Pacific, and around the world, having widespread climate and socio-economic impacts. For example, in the past El Niño was blamed for forest fires that created dangerously polluted air in Indonesia, water shortages from drought in northern Brazil or massive fish migration from the warmed waters along the coasts of Peru and Ecuador. These had devastating economic consequences. However, in some cases El Niño conditions appear to lead to benefits. For example, the phenomenon is associated with warmer winters in the northern US, resulting in reduced energy use and therefore, lower energy prices [23].

6.1.4 Predicting El Niño

El Niño is linked with impacts on many aspects of human life. It affects commodity prices and the macroeconomy of many countries [20]. It appears to account for changes in global crop production [52], price inflation and world economic activity.

The phenomenon is associated with health impacts [62], such as increased risks of malaria epidemics in parts of South Asia and South America or civil conflicts [51].

The number of climate disasters does not change between years with El Niño conditions and neutral periods. However, the types of catastrophe to be expected are more easily predicted in El Niño years allowing preparation for them [46]. Being able to better foresee the phenomenon would help water, energy, and transportation managers, and farmers, plan for, avoid or alleviate potential losses. As presented earlier in this section, El Niño affects the agriculture, fishing, forestry and energy sectors, as well as commodity prices. Improved climate predictions could have an impact in these areas.

It would be of great value to humanity to be able to predict the phenomenon in the Pacific. This motivates us to study predictability and makes it an important statistical issue. We now move to mathematical details of dynamical systems and predictability in the context of the Cane-Zebiak model, which is used to forecast El Niño events.

6.2 Cane-Zebiak model

6.2.1 Information about the model

In the early 1980s, scientists based at Columbia University's Lamont-Doherty Earth Observatory constructed the first oceanic-atmospheric model, which aims to predict an El Niño event [1]. The model computes anomalies of atmospheric and oceanic fields, relative to a specified monthly mean climatology that is calculated from past observations, for the tropical Pacific region [21].

The C-Z model is a deterministic numerical model of the coupled ocean and atmosphere, which is used to study ocean-atmosphere interactions in the tropics. C-Z model produces recurring El-Niño events, irregular in both amplitude and spacing [114]. See also Fig.6.6. They occur around every 3-4 years, with the largest growth of Sea Surface Temperature during summer and autumn, lasting around a year, which is in general agreement with the main features of observed El Niño events. No statistical procedures are used in the forecasts [21]. Model variables evolve deterministically according to physical principles thought to govern the atmosphere and oceans. We do not review the physical laws used in the model's equations; these are presented in a paper 'A Model El Niño-Southern Oscillation' [114] written by the authors of the model.

The C-Z model is written in Fortran. The model consists of fifteen different variables. We restrict our interests to one - Sea Surface Temperature (SST). Later in this chapter we analyse the Niño 34 index, which is the average SST anomaly within the region 5°S–5°N, 170°E–120°W. Earlier in section 6.1.1 we introduced the Niño regions and Niño 34 index. Niño regions used by the C-Z model are illustrated in Fig.6.1.

Some model parameters are adjustable, but the integration time step and locations of each grid box are fixed somewhere in the program. The default value of a time step is 10 days. In the research we focus mostly on the Perfect Model Scenario. We also consider an Imperfect Model Scenario, which is the C-Z model with a 5-day time step as the model.

The standard grid for the SST physics and atmosphere model is 5.625° longitude by 2° latitude. The model simulates the results for the area between 101.25°E to 95.62°W and 29°N to 29°S. In C-Z model the area of forecast is arranged as an array,

which has index I corresponding to longitude running from west to east, and the index J corresponding to latitude and running from south to north. I=1 corresponds to 101.25°E, with successive points 5.625° degrees apart. J=1 corresponds to 29°N, with successive points 2° degrees apart. The overall domain in C-Z model consists of 1020 grid points - 30 different latitude boxes and 34 different longitude boxes.

This time, building the components of the research and running the simulation has been more complicated and time consuming than in low-dimensional mathematical model cases. (I am grateful for receiving training on this from colleagues from the International Research Institute for Climate and Society (IRI) at Columbia University in New York). For example, for the experiment described in section 6.3, the input size is around 3.7 GB and the time required to do the simulation is approximately 16 hours. Input consists of 2^{14} ICs, each IC is represented by a file consisting of SST anomalies from the mean, for the total area of forecast in the model, that is for 1020 grid points. Running the model and simulation requires a Linux operating system.

6.2.2 Qualitative Behaviour

SST in the Niño 3.4 region

Fig.6.6 shows 64 years of output from the Cane-Zebiak model (C-Z model). The red and blue lines illustrate the evolution of mean monthly sea surface temperature anomalies in the Niño 3.4 region of the eastern tropical Pacific (also known as NINO34 index) from two different initial conditions. For comparison, the black dotted line shows observations between 1950-2014. It appears that the C-Z model is able to simulate important aspects of interannual variations in the tropical Pacific realistically. The time series of the model are aperiodic, with decades of irregular

cycles of warm phases with an average period of 4 years.

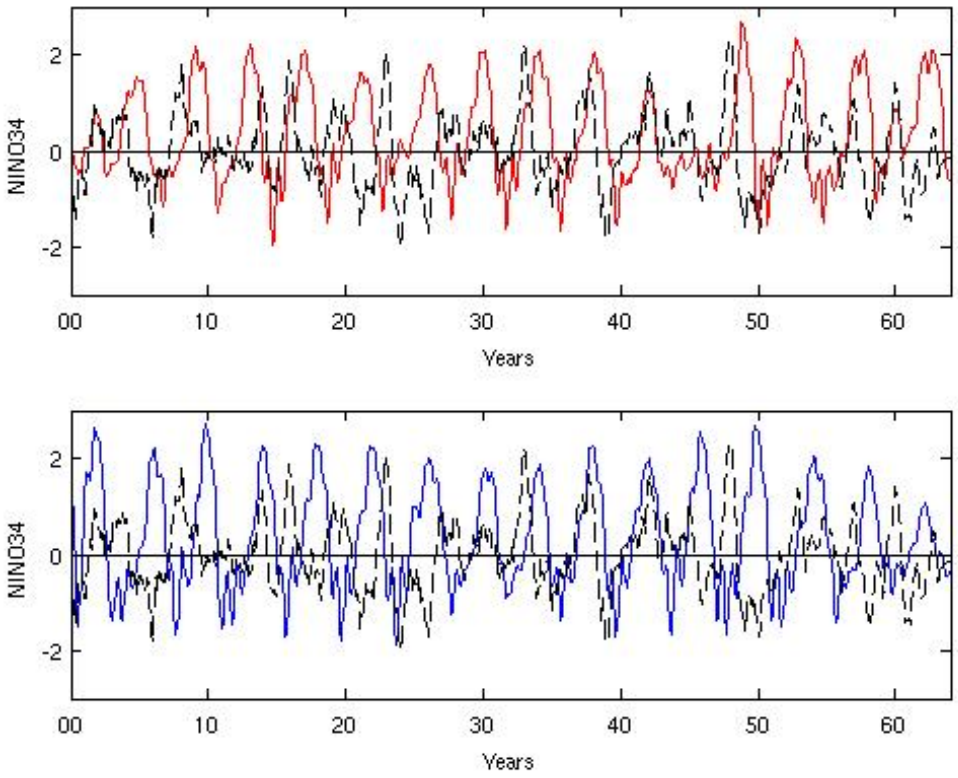


Figure 6.6: Examples of time series of the monthly NINO34 index for 64 years from the C-Z model (red and blue lines) vs. 1950-2014 observations by NOAA (black dotted line)[7]. Initial conditions in the top and the bottom charts are different. In both cases we observe aperiodic cycles of warm phases with an average occurrence period of 4 years. The range of the index is between -2 and 3°C. The model appears to illustrate fluctuations in SST in the tropical Pacific reliably (see section 6.1.1 for more details). This behaviour is similar to observations, and the model looks non-linear.

For the Perfect Model, the NINO34 index stays within the range between -2 and 3°C. This is in contrast to output from the Imperfect Model (IM). Fig.6.7 shows a time series of the 5-day model versus the time series with exactly the same initial conditions for the 10-day model. Here we notice that the time series looks very different after 4.5 years. A discrepancy of 0.5°C (magnitude of El Niño threshold)

is seen at that point. By contrast, after 1.5 years the difference is 0.1°C.

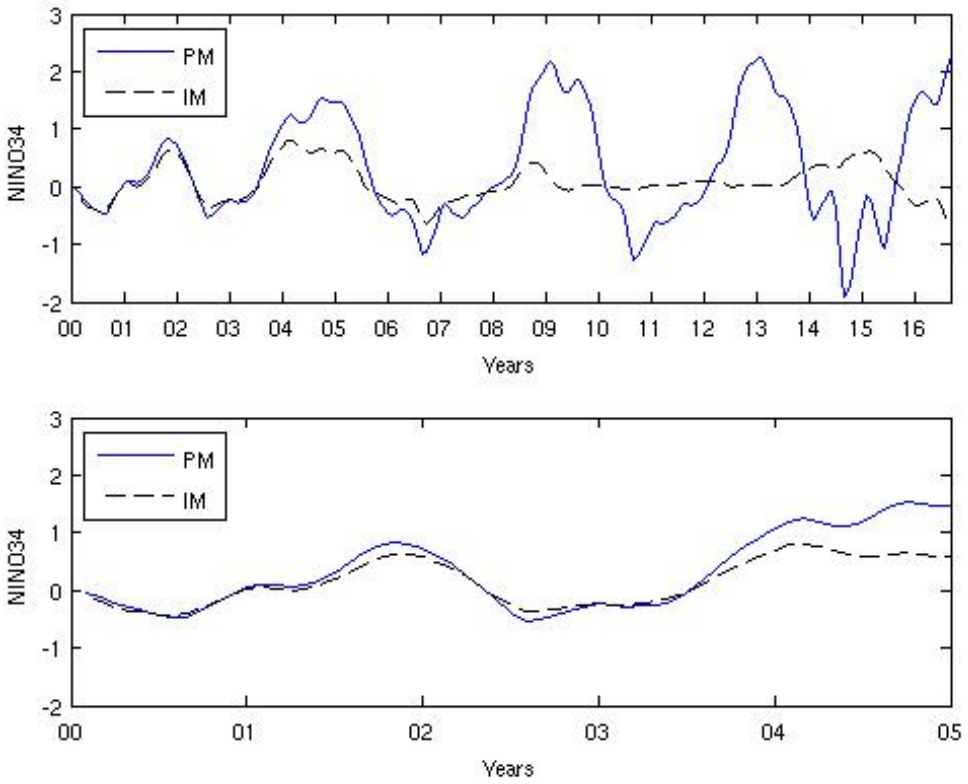


Figure 6.7: Trajectories of the 10-day model (used in the Perfect Model Scenario) and the 5-day Model (used in the Imperfect Model Scenario) of the C-Z model. Time series of the Perfect Model (solid blue line) is evolved from exactly the same initial condition as the Imperfect Model (dotted black line). The top chart illustrates the first 16 years, while the bottom one zooms into the first 5 years. After 1.5 years, both model types differ by 0.1°C. After 4.5 years they diverge at the El Niño threshold level of 0.5°C. They start to look very different at that time.

Nonlinear behaviour

We introduced the theory of chaos earlier in chapter 2.3. Although we do not attempt to show that the C-Z model is chaotic, in Fig.6.8 we demonstrate sensitivity to initial condition error in the C-Z model. The green line shows the target forecast; this is a time series of the NINO34 index. We introduce small changes to the initial

condition. This results in large differences between the trajectories later in time. Time series representing the smallest initial separation (red line) stays the closest to the true trajectory the longest. The trajectory from the biggest noise level (blue line) diverges the earliest of the three considered. The fact that the trajectories eventually diverge, demonstrates that the C-Z model is sensitive to initial condition error.

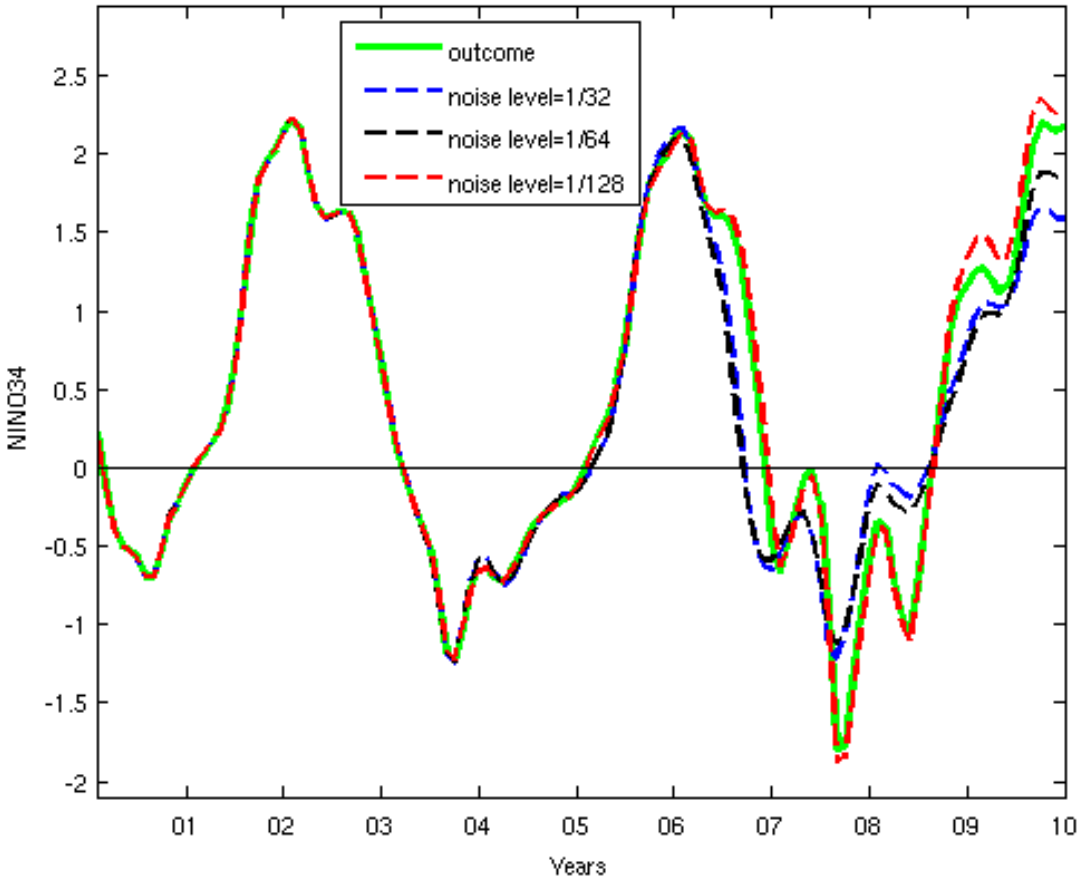


Figure 6.8: Trajectories of the NINO34 index of the C-Z model with 3 different noise levels. Nearby initial conditions separate from the target forecast. At year 10 we observe that the time series representing the smallest noise level of 1/128 (red dotted line) stays the closest to the outcome (green solid line). The biggest noise level of 1/32 (blue dotted line) is the furthest from the target forecast.

The analysis in this section shows that the C-Z model appears to embody the qualitative features of El Niño in the tropical Pacific. NINO34 index trajectories demonstrate that there is sensitivity to initial conditions. This suggests that Columbia Universitys model for El Niño is suitable for the application of non-linear analysis methods, which I developed earlier in this thesis.

6.3 Decay of information in the Cane-Zebiak model - Perfect Model Scenario

In the previous chapter, we demonstrated the decay using the Logistic Map as a model of the Quartic Map. Here, again, we use the Ignorance score to demonstrate for how long the average time for which model is informative. In addition we calculate the Information Deficit and demonstrate how it can be used to find weaknesses in the forecasting system (including bugs in the code). In this section, we focus on the Perfect Model Scenario, where we use one form of the C-Z model as another form of the C-Z model. First, in section 6.3.1, we describe the experimental design. In section 6.3.2, we explore the behaviour of the climatology. We then determine the operational parameters of the kernel dressing used to build forecast densities. In section 6.3.4, results of the experiment are described and interpreted.

6.3.1 Generation of ensemble-based probabilistic forecasts of the NINO34 Index

The concept of ensemble-based probabilistic forecasting (Monte-Carlo sample) was introduced in section 2.6. The experiment on tracking the decay of information in the C-Z model involves creating inverse noise ensembles, evolving them forward and

extracting time series of the target values. Our variable of interest is the sea surface temperature anomaly in the Niño 3.4 region which defines the Niño 3.4 index (or NINO34 index). Both Niño 3.4 region and Niño 3.4 index were introduced earlier in section 6.1.1. In C-Z model domain, the Niño 3.4 region is between latitude fields (J) from 16 to 26 and longitude fields (I) from 13 to 18, the information about a grid point in C-Z model is in section 6.2.1.

In the simulation we perturb the initial condition of the SST field only, by picking up 6×11 consecutive grid points from the domain of 12°N to 1°S and 130°E to 115°W. That size of initial condition represents size of Niño 3.4 region in the model. The location of the perturbed initial condition is close to the area defined by the Niño 3.4 region in the real world 5°N to 5°S and 170°E to 120°W.

First the input for the simulation is created. Details of the experiment, which we call experiment 6.A, are listed in Table 6.1 and described below.

Variable to perturb	SST (Sea Surface Temperature)
Number of grid points (total)	1020 (30×34)
Number of grid points (IC)	66 (11×6)
Index grid point (lon min)	13
Index grid point (lat min)	6
Mean SST	2
Noise level	0.01
Perturbation	U (1,3)
n_{ens}	256
m	64

Table 6.1: Experimental Design 6.A

Assume we observe 66 grid points of SST anomaly from the mean, this is drawn from 1020 grid points, and x_i for i=1,...,66 are system variable on U(1,3), (as this value could simulate the ocean warming near Niño 3.4 region). All other variables are set to their climatological mean.

We propagate this IC forward to generate system trajectory.

To generate an IC ensemble $EnsIC_{i,j}$ we add Gaussian noise as

$EnsIC_{i,j} = x_{0,i} + \varepsilon_{i,j}$, $\varepsilon_{i,j} \overset{iid}{\sim} N(0, 0.01^2)$

For i=1,...,66 and j=1,...,m

where $\varepsilon_{i,j}$ is a vector of iid draws from a Gaussian distribution with mean 0°C and standard deviation $S_d = 0.01$°C.

The input consists of $n_{ens} = 256$ ensembles, each consisting of $m = 64$ members. After this step, each ensemble with perturbations is evolved forward in time using the C-Z model. Details of experiment 6.B are given in Table 6.2. The lead time unit is one month and the final output includes 64 years of the NINO34 index. During this experiment climatology is also built, which is described in the next section.

System	C-Z model (integration time step 10 days)
Model	C-Z model (integration time step 10 days)
n_{ens}	256
Noise level	0.01
m	64
Lead time unit	a month
ts (model)	768 (64 years $\times$ 12 months)
ts (system)	768 (64 years $\times$ 12 months)
n_{icclim}	8
$TRAN_{clim}$	1200 (100 years $\times$ 12 months)
ts (climatology)	22800 (1900 years $\times$ 12 months)
Number of climatologies	12 (for each month)
Size of climatology	15200 (1900 years$\times$ 8 IC)

Table 6.2: Experimental Design 6.B

Fig.6.9 shows an example of a Monte-Carlo sample developed with the C-Z model for a period of 10 years. Most members stay close to the target forecast for the first 2 years and at years 3.5 and 5.5. That is when the forecast is good. After 10 years, nearby initial conditions spread along the attractor.

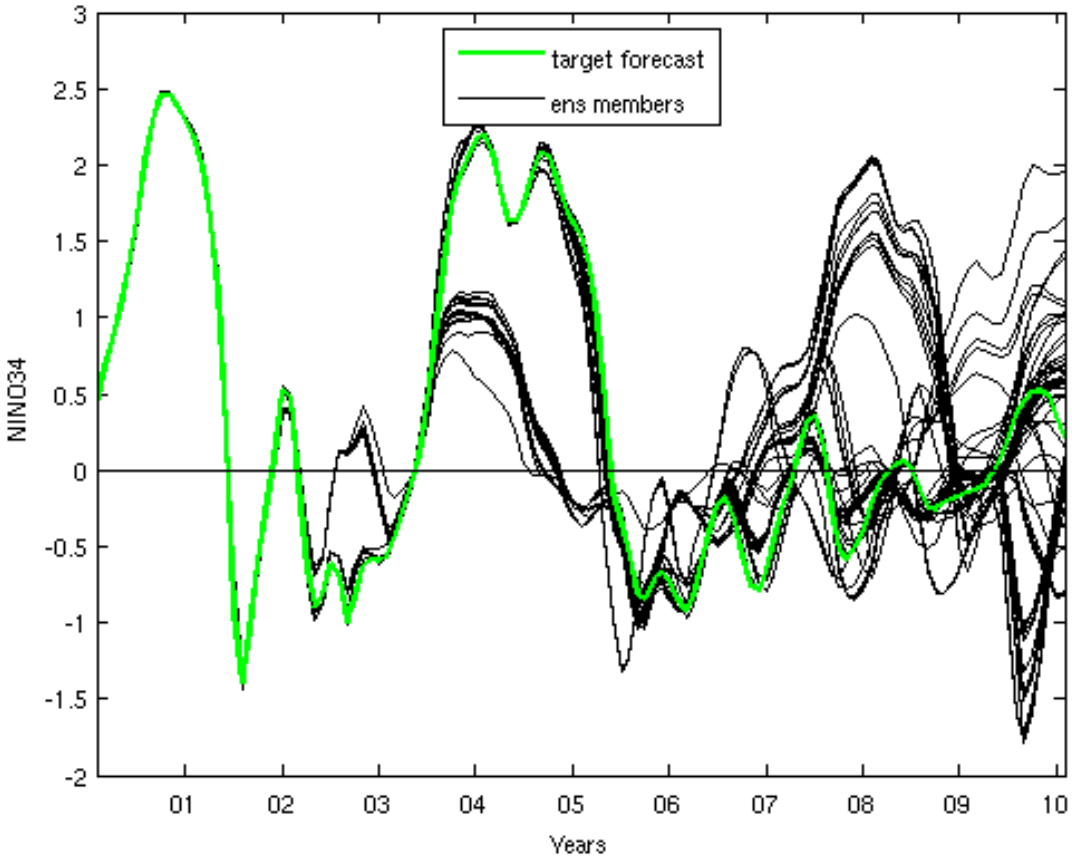


Figure 6.9: Example of an ensemble forecast of the NINO34 index evolved on the C-Z model in experiment 6.B. There are 64 scenarios shown with black lines. The green line is the outcome. Most members stay close to the target forecast for the first 2 years, that is when the forecast is good. Then at year 3, if we turned nearby initial conditions into a probability forecast it would be a bimodal pdf, because members of the ensemble oscillate around two values. At year 3.5 there is a return to a skill, when everything is back in the same place again. At years 4 and 5 we have something bimodal, in year 5.5 all members are together again and we have a good forecast. Then nearby initial conditions get separated along the attractor. This behaviour would not be the same for every ensemble. This example illustrates a sensitivity to initial condition.

6.3.2 Climatology for NINO34 for the C-Z model

Obtaining a climatology for NINO34 from the C-Z model

Climatology was introduced in chapter 2.2. To derive a representative climatology of the model, ideally, we would hope to be able to draw randomly from the model attractor. We take $n_{icclim} = 8$ initial conditions and evolve each of them forward $TRAN_{clim} = 100$ years, such that each trajectory has had time to reach non-transient state. We then evolve each of these states a further 1900 years in time and sample each trajectory and use these points as our climatology. We recognise that this approach is suboptimal in terms of selecting random points from the climatology, but we expect this to give a reasonable representation. This was the approach recommended by scientists from IRI at Columbia University in New York. (We will refer to this set as **Clim1**).

Comparing 2048 Gaussian preturbations to Clim1 (experiment)

We compare Clim1 to an alternative method of creating a climatology, where the input is n_{ic} (2048) Gaussian perturbations around one initial condition with standard deviation S_d of 0.01°C, evolved for 64 years in a specified month - May. We will refer to this set as **Clim2**. Clim2 defines a model, while the Clim1 defines the system. We will refer to this experiment as Experiment 6.C.

System	C-Z model (integration time step 10 days)
Model	C-Z model (integration time step 10 days)
Time unit	a month
Selected month	May
Noise level	0.01
ts (system)	1900 (years)
ts (model)	64 (years)
n_{ic} (system)	8
n_{ic} (model)	2048
$TRAN_{system}$	100 (years)
Number of bins (RE)	8

Table 6.3: Experimental Design 6.C

To measure the difference between the two climatologies we use Relative Entropy (RE) computed in two different ways: RE with equally likely bins and RE with equally spaced bins, as discussed in section 3.5.

We now look at RE between 2048 Gaussian perturbations around one IC (model) and Clim1 (system) defined earlier, for equally likely and equally spaced bins (see Fig.6.10). The number of bins considered here is 8. RE approaches equilibrium after 50 years (but does not approach exactly zero). Thus distributions of model and system become similar after around 50 years and the climatological set defined as Clim1 is right. Perhaps evolving points for only 50 years would not be enough. Another observation that we make here, is that the RE curve in Fig.6.10(a) is flatter than in Fig.6.10(b). In case (b) fluctuations of bins with fewer number of members

would be bigger and RE is bigger. If all the bins have the same probability, then variations are smaller and RE is smaller. Bins with fewer number of members in bins for the system have low probability, so we divide by a small number and get a large RE. We also notice here the zigzag shape of the RE curves. This reflects the fact, that the NINO34 index is at its highest or lowest peak every two years on average. When, the forecast is in states of climatology that are more likely to happen, the RE is lower.

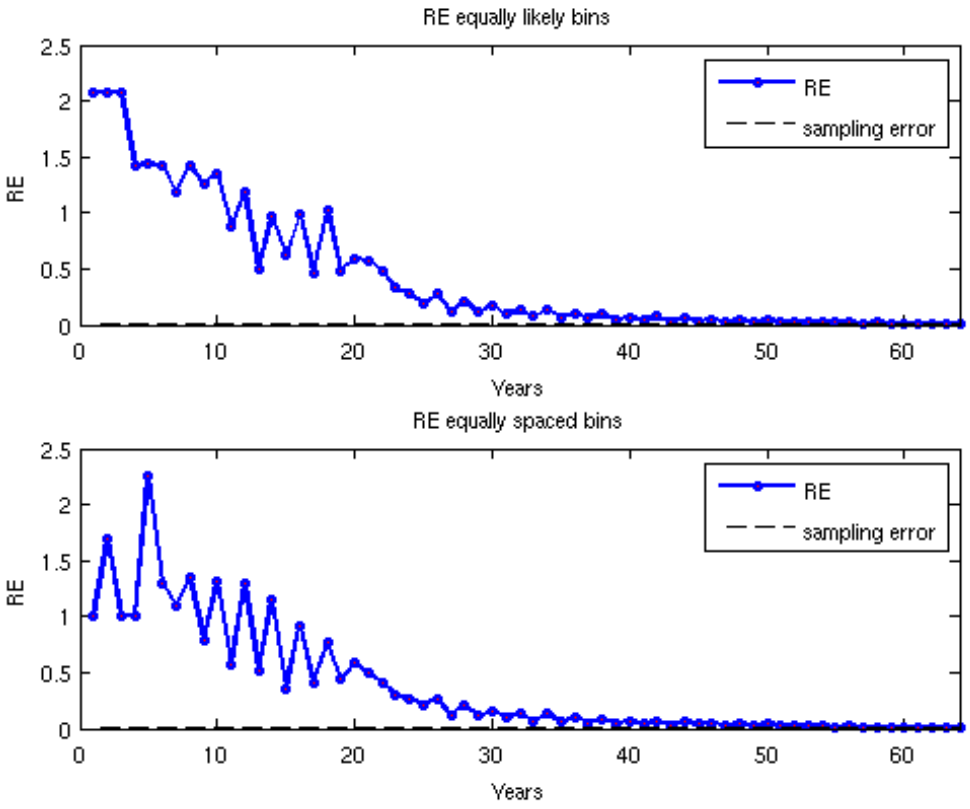


Figure 6.10: Relative Entropy (RE) between two distributions of Clim1 and Clim2 in experiment 6.C. This is plotted for 64 years in May for (a -top chart) a case with equally likely and (b -bottom chart) equally spaced bins. On a scale of 50 years RE approaches the sampling error line, which means that probability distributions representing both methods look similar after that time. We notice here, that the shape of the RE curve is spiky. This happens when variation of bins with fewer members is big, so the RE is big for those bins. In the cases of low state or high state of the NINO34 index we have poor skill (high RE), in between (in the middle) the skill is good and RE smaller, because that is what the climatology looks like. Good skill happens every two years, which explains the zigzag shape of the RE curve. In case (a) fluctuations are smaller and so the curve is less spiky than in case (b).

Fig.6.11 shows distribution of the probability of model (Clim2) relative to system (Clim1) by bin and year for May. For every year we colour each box representing the bin, by the probability of a Clim2 being in these bins. Black colour means

probability equals one, where all elements are in the box. Light yellow indicates less likely bins. Each vertical slice adds up to one. Again, we consider two cases here: (a) equally likely and (b) equally spaced bins. At the beginning of the simulation all members are in the same box (because they were derived as many Gaussian perturbations around one initial condition). They spread around all bins quicker in case (a) than (b). At the end of the simulation in year 64, all boxes are light yellow in Fig.6.11(a), as all bins are equally likely. This is different in Fig.6.11(b), where box 4 is the most probable after 60 years. If we drew a frequency distribution at year 64, its shape, as expected, would be flat in case (a); in case (b) it would look like a climatology. The case with fixed bins has more noise than the one with equally likely bins. The presence of orange stripes in almost every box means the model did not converge to the system, and the information has not decayed here.

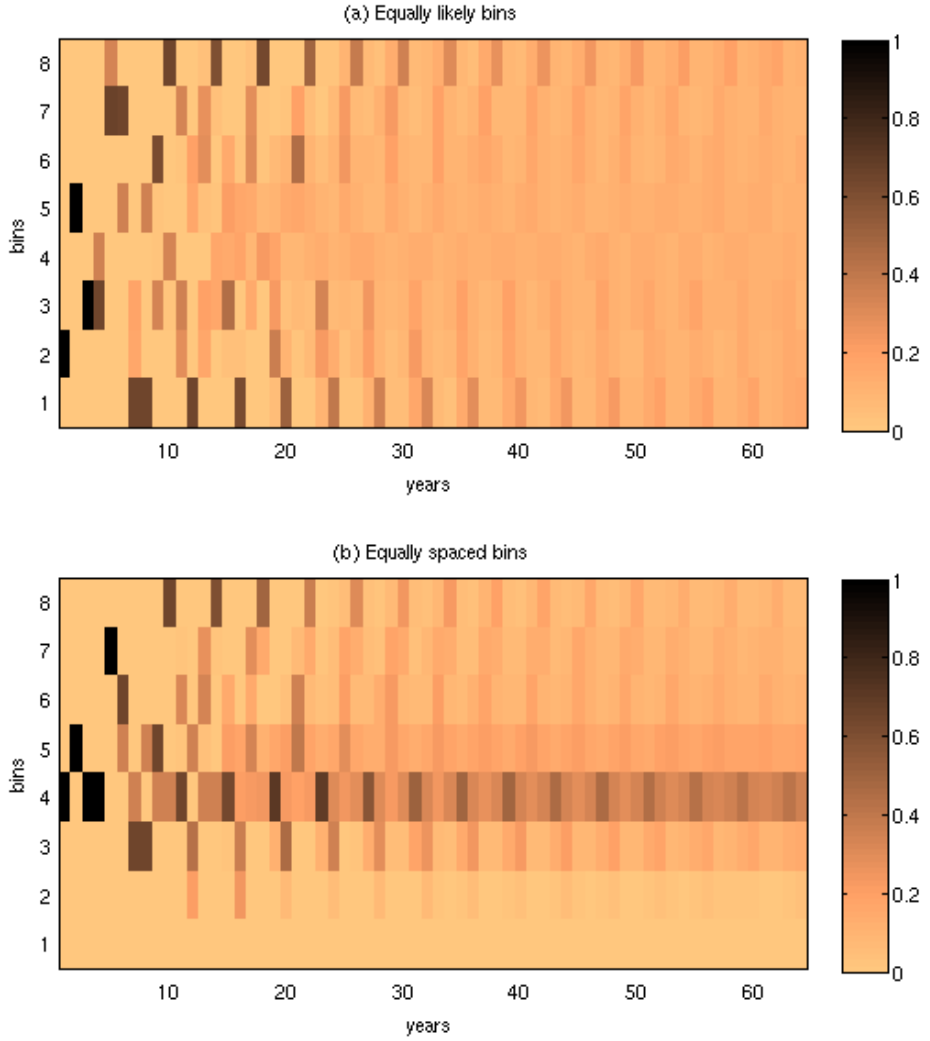


Figure 6.11: Proportion of 2048 Gaussian perturbations around 1 IC (Clim2) relative to a Clim1 in time and by bin, experiment 6.C. In fig.(a) equally likely, in (b) equally spaced bins are shown. Colour relates to the proportion of model elements in that box. Black means the most likely, light yellow indicates the least probable. Each vertical slice adds up to one. At the beginning all elements are in the same box. Here we start with all light yellow and one black box, which becomes lighter with a time. Colour in boxes flickers with a time step. In case (a) with equally probable bins, each box is the same at the end, because all elements are equally spread around each bin. In case (b), box 4 is the most likely after 60 years. Fig.(a) decays flat and fig (b) decays to climatology. Gaussian perturbations spread around all bins quicker in case (a) than (b). There are still oscillations in almost every box (see orange stripes across every bin), which suggests that the information has not been lost.

Now, we find an equilibrium (sampling error) to which RE approaches. This is illustrated with a dotted black line in Fig.6.10. Details of the experiment, which we call experiment 6.D, are listed in Table 6.4. We draw n_{ic} number of uniformly distributed random numbers between 0 and 1, then divide them into k number of bins and calculate the number of elements in each bin. We repeat this many times (2^{20}, around million) and estimate the average RE for n_{ic} number of elements and k number of bins. Two cases are considered: equally and unequally likely bins. Here an equally spaced grid means that the bins have an equal probability of being drawn. Unequally likely bins have intervals on the grid determined by the vector of different fractions.

n_{ic}	where $n_{ic} = \{128; 256; 512; 1024; 2048\}$
IC	U(0,1)
k (number of bins)	where $k = \{4; 8; 16; 32\}$
Number of draws (repetitions)	2^{20}

Table 6.4: Experimental Design 6.D

Results for n_{ic} and k for both methods are in Table 6.5 and Table 6.6. Here we notice that RE decreases proportionally to the increase in the size of the sample (n_{ic}). Also, the bigger the number of bins (k) the bigger the RE. The difference between equally and unequally likely bins case is shown in Tab.6.7. For most (n_{ic},k) combinations RE is bigger for the unequally likely bins method.

		Size of the sample (n_{ic})				
		128	256	512	1024	2048
Number of bins (k)	4	0.0118	0.0059	0.0029	0.0015	0.0007
	8	0.0277	0.0138	0.0069	0.0034	0.0017
	16	0.0601	0.0296	0.0147	0.0073	0.0037
	32	0.1291	0.0621	0.0306	0.0152	0.0076

Table 6.5: Relative Entropy by size of the sample (n_{ic}) and number of bins (k) for equally likely bins case in experiment 6.D. Data was drawn from the uniform distribution 2^{20} times, values in the table show RE averaged over the number of repetitions. The bigger the number of bins (k) the bigger the RE. It decreases proportionally to the increase in the size of the sample (n_{ic}).

		Size of the sample (n_{ic})				
		128	256	512	1024	2048
Number of bins (k)	4	0.0119	0.0059	0.0031	0.0015	0.0007
	8	0.0275	0.0141	0.0069	0.0035	0.0017
	16	0.0624	0.0301	0.0150	0.0075	0.0037
	32	0.1269	0.0625	0.0303	0.0154	0.0077

Table 6.6: Relative Entropy by size of the sample (n_{ic}) and number of bins (k) for unequally likely bins case in experiment 6.D. Data was drawn from the uniform distribution 2^{20} times, values in the table show RE averaged over the number of repetitions. Here conclusions are the same as in Tab.6.5. The bigger the number of bins (k) the bigger the RE. It decreases proportionally to the increase in the size of the sample (n_{ic}).

		Size of the sample (n_{ic})				
		128	256	512	1024	2048
Number of bins (k)	4	-0.0001	-0.0001	-0.0001	0.0000	0.0000
	8	0.0002	-0.0003	-0.0001	0.0000	0.0000
	16	-0.0023	-0.0005	-0.0003	-0.0001	0.0000
	32	0.0022	-0.0004	0.0004	-0.0002	-0.0001

Table 6.7: Difference in Relative Entropy by size of the sample (n_{ic}) and number of bins (k) between 2 methods: equally and unequally likely bins in experiment 6.D. Data was drawn from the uniform distribution 2^{20} times, values in the table show RE averaged over the number of repetitions. There is not much of a difference between the methods. In most of the cases RE for unequally likely bins is bigger.

RE shown in Fig.6.10 is above the sampling error with 95% resampling intervals, even after 60 years (see the zoom in Fig.6.12). This means that the probability distributions of model and system are still different and perhaps the data in this experiment is too short. As the number of time steps in the experiment increases any pattern in bins would be lost and the RE curve will approach the sampling error line.

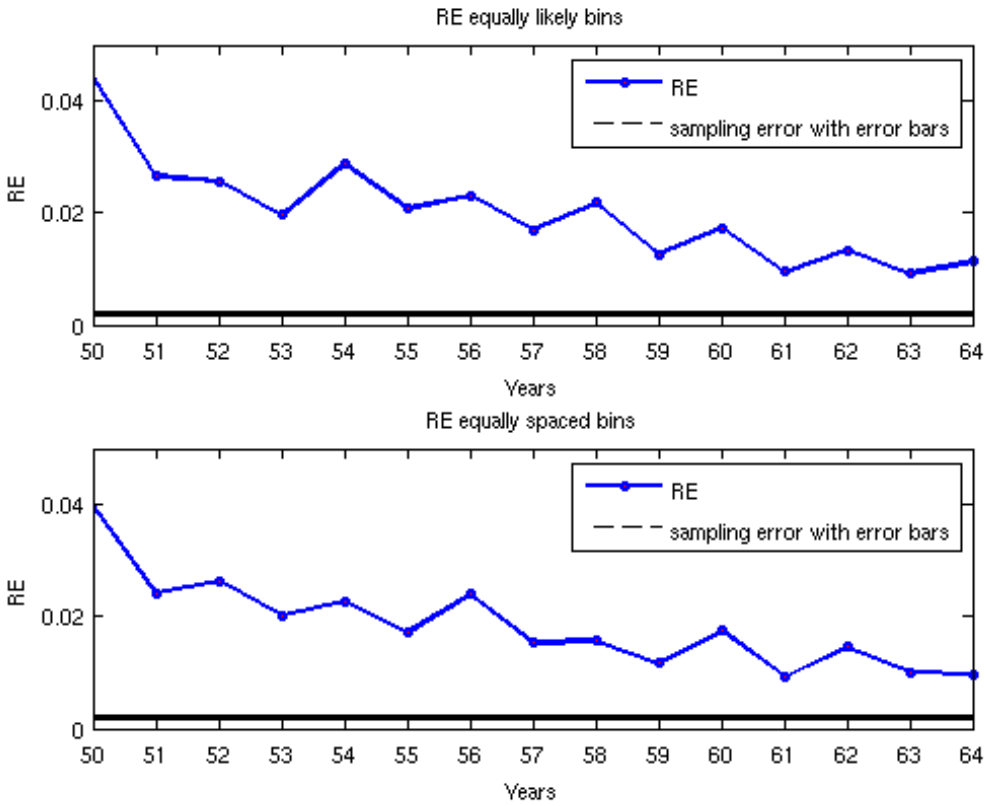


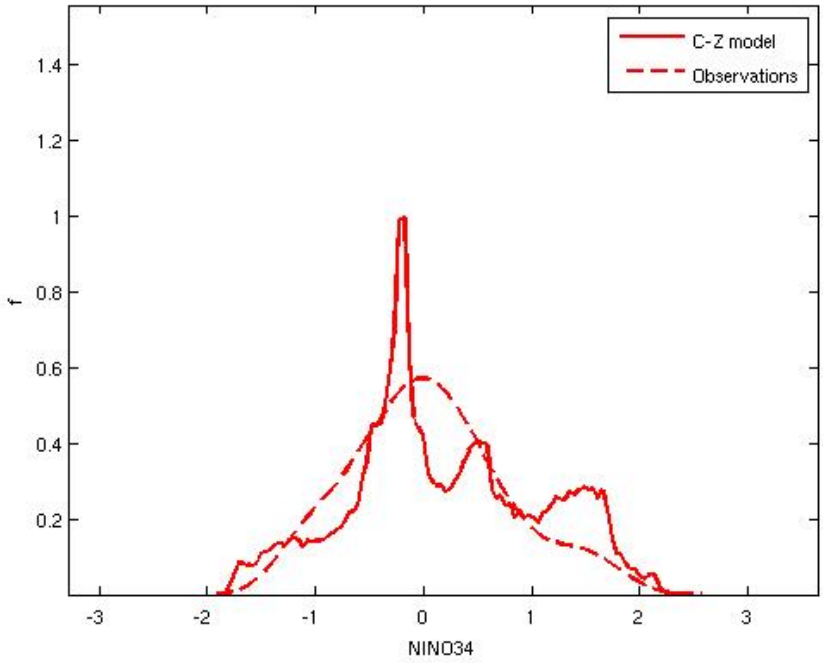
Figure 6.12: Tail of the *RE* curve shown in Fig.6.10. Here *RE* is above the sampling error (calculated in Tables 6.5 or 6.6) with 95% resampling intervals for both (a) equally likely and (b) equally spaced bins. This could mean that the data used in the experiment 6.C is too short for the model to converge to the system, and could explain the pattern across bins (orange stripes) in Fig.6.11, still visible after 60 years.

Climatology C-Z model vs. observations from the real world observations?

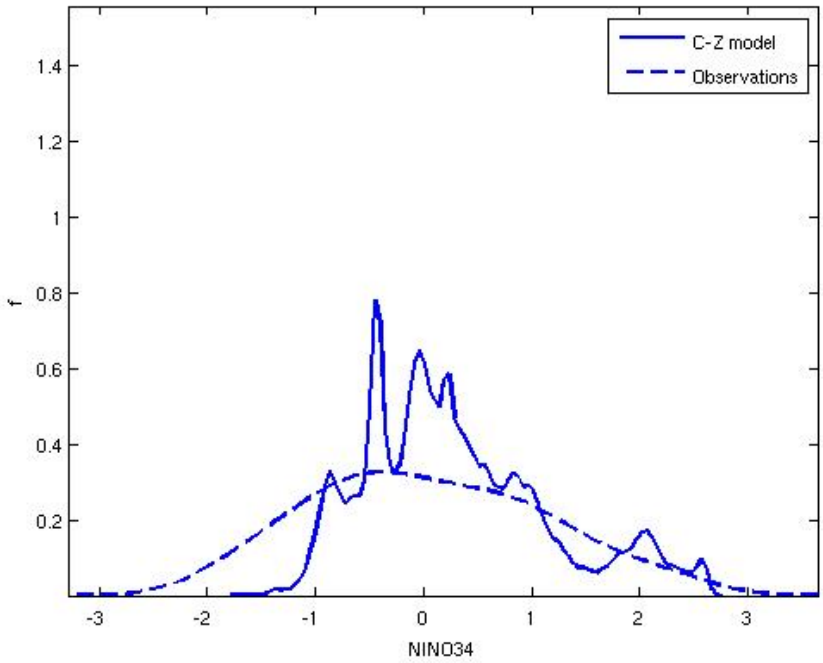
In chapter 5 we discussed the decay of predictability and showed the climatology of a simple dynamical system (Logistic Map). One new aspect of the C-Z model case, is that the model climate for each month can be estimated, so there are 12 different climatological states. In addition to this new feature, model climate can

be compared to climatology estimated using 1950-2016 observations.

Fig.6.13 shows climatology for both model (solid line) and observations (dotted line). These are typical values one can expect by running the model for a long time (in this case 8 initial conditions for 1900 years) or by looking at old statistics (66 years of observations). The red curve in Fig.6.13 (a) illustrates the calculated probability density estimate of climatology for the NINO34 index in August, the blue one in Fig.6.13 (b) represents December. The shapes of the lines look very different between months. In August, the model tends to be in one state, somewhere around 0°C of a difference from the mean, while in December there is much more of a spread of different possibilities in different years. This broadly agrees with observations that in December the NINO34 index at its extremes, while in August it decays (this is expressed by the narrower range of values on the X-axis in case (a), which gets wider in case (b)).



(a) August



(b) December

Figure 6.13: Climatology of NINO34 index for C-Z model (solid line) and 1950-2016 observations by NOAA (dotted line). Shapes of probability density estimate between (a) August and (b) December are very different. In case (a) the model is around one state - 0°C, while in case (b) it is spread around different possibilities. This corresponds to the observations, as the NINO34 index reaches its extremes in winter and diminishes in summer. In case (b) the model does not capture the entire range of values.

6.3.3 Adjusting parameters of kernel dressing of Niño 3.4 Index ensemble forecasts

In this section, we use the ensemble and climatology described in experiment 6.B to build probability forecasts. As before, we use kernel dressing, optimising the parameters over the ensemble-outcome archive, as introduced earlier in chapter 3, in section 3.6. A different set of parameters are found for each month.

Fig.(6.14) shows the kernel width σ and the weight α for four chosen months (Jan, Apr, Jul, Oct). In general, the value of σ increases very slowly over the first 10 years. The value of α is close to 1 at short lead times and tends to decrease. This decrease is far less smooth than in the examples presented in chapter 5. When α is large, the ensembles contribute significantly to the forecast and can be considered useful.

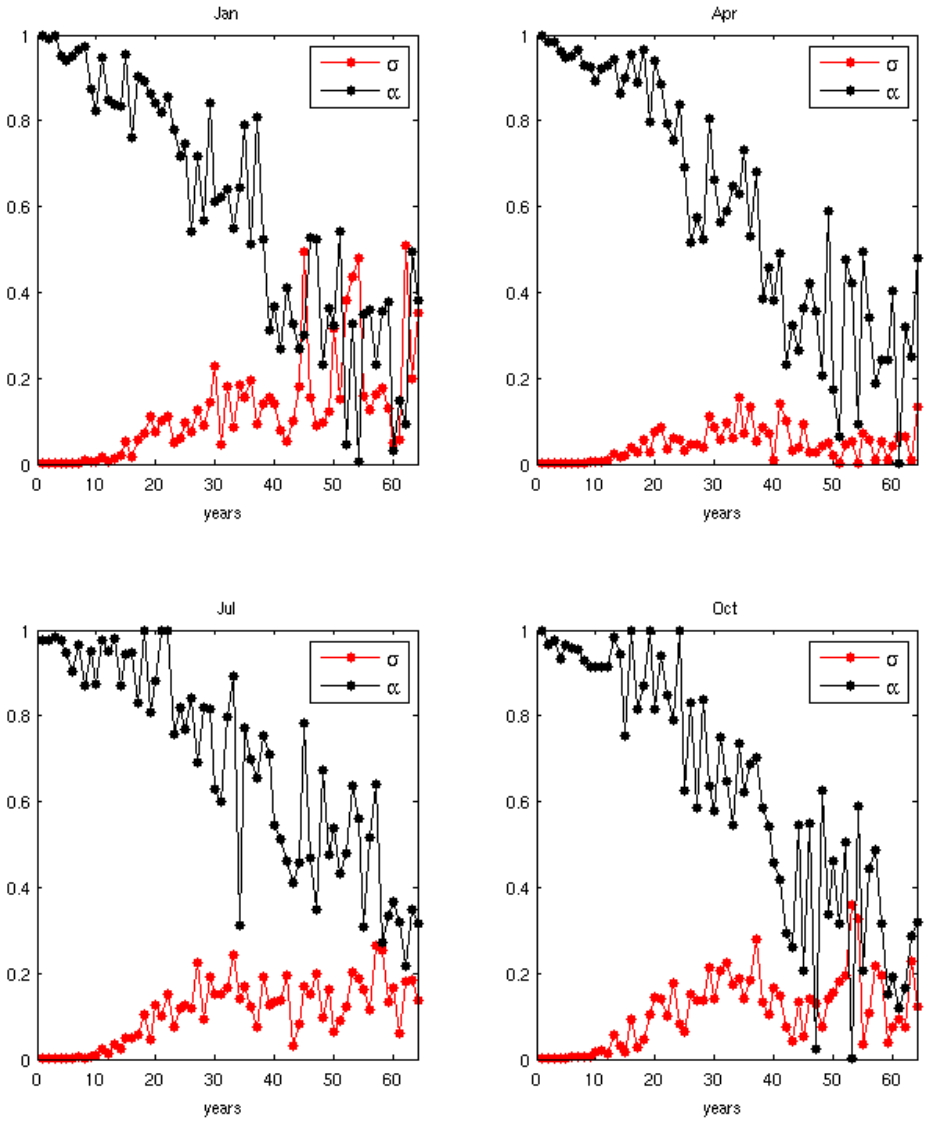


Figure 6.14: Operational kernel dressing parameters: the kernel width σ (red line) and the weight α (black line) for four chosen months as a function of lead time in experiment 6.B. Although α with lead time tends to decrease there is a wide variation in this parameter value over time. The value of σ is also rather volatile.

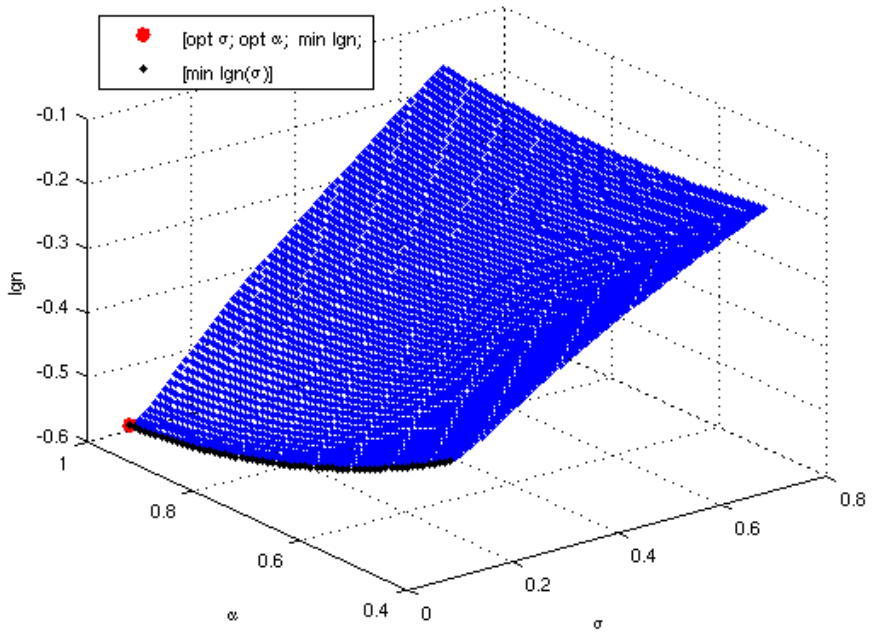
Next we analyse the effect of changing the parameters (α, σ) on the mean Ignorance.

One would expect α to decrease with time. In Fig.6.14 we see that this is not always

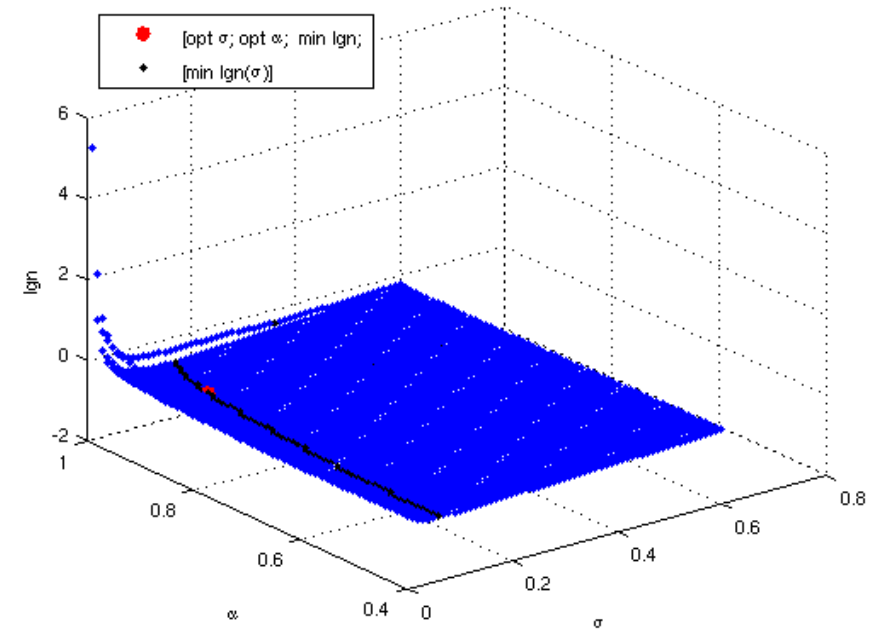
the case. For example, if we look at α in October in year 24, we notice that it is higher than in previous steps and it is equal to 1. In the Logistic Map α decays with lead time, as shown in Fig.5.5. In the C-Z model in October, in year 24 α = 1, σ = 0.082 and mean Ign = -0.59. The mean Ignorance as a function of α and σ over the in-sample and out-of-sample sets are shown in Fig.6.15. In both cases, the red dot shows the point at which the mean Ignorance is minimised. In case (b), the mean Ignorance with slightly different parameter values as is expected through sample variability. Results are shown in Table 6.8. The black line shows the minimum Ignorance for every value of α.

	in-sample set	out-of-sample set
α	1.00	0.89
σ	0.082	0.12
Ign	-0.59	-0.58

Table 6.8: Operational parameters and Ignorance in October in year 24 derived from the in-sample and out-of-sample sets in experiment 6.B. As expected, minimum Ignorance and blending parameters between sets are different. These results are the coordinates of the red dot in Figure 6.15 (a) and (b).



(a) in-sample



(b) out-of-sample

Figure 6.15: Ignorance by weight (α) and kernel width (σ) at year 24 in Oct. The mean ignorance of the in-sample set (top) and out-of-sample set (bottom) at year 24 in October, as a function of α and σ. The red dot shows the point at which the mean ignorance is minimised in each case. The black line shows the lowest value on this surface. This demonstrates how the optimised parameters can be sensitive to the sample used.

Fig.6.16 is a contour plot of α vs. σ, for the in-sample set. The plot is shown to demonstrate the relationship between the parameters. Labels on the contour lines show values of the mean Ignorance. When α increases and σ decreases the mean Ignorance tends to decrease. This shows that, as we move to bigger α the kernel is slightly increased to compensate.

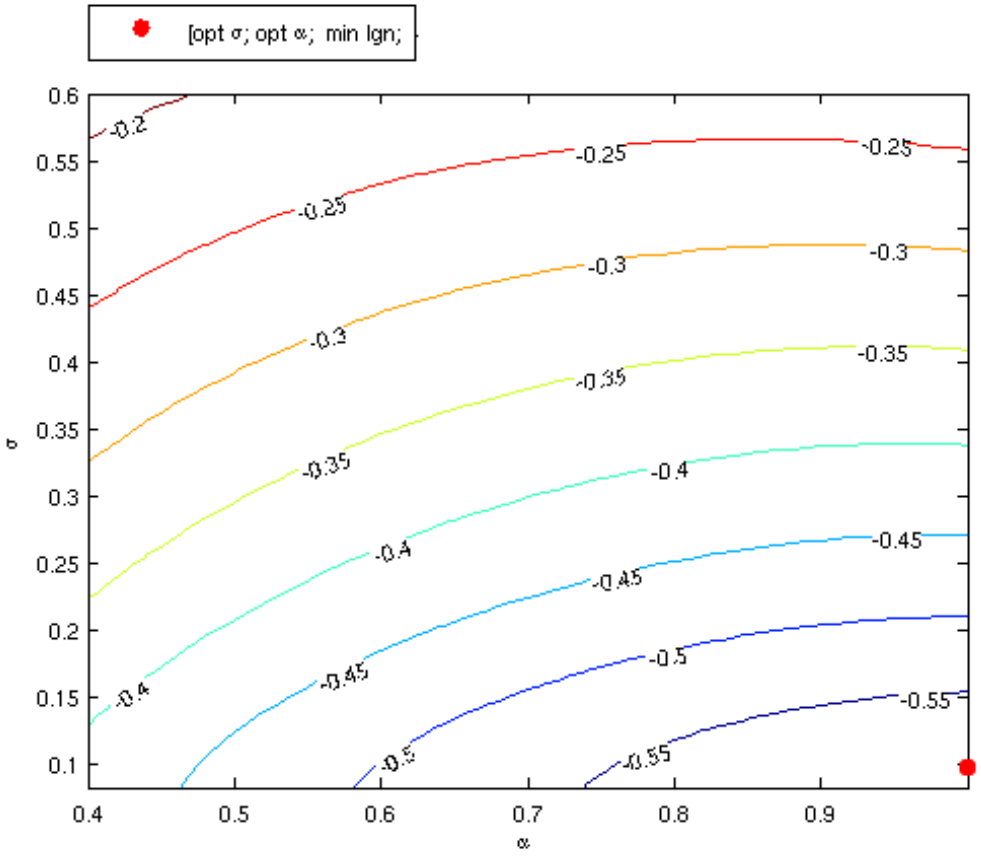


Figure 6.16: Contour plot of mean Ignorance with α and σ. The Mean Ignorance tends to decrease when α increases and σ decreases. Here, we also notice that as we move to bigger α, the optimal value of σ tends to increase. The red dot shows combination of α and σ that minimises the mean Ignorance as shown in Fig.6.15.

In order to understand the operational parameters shown in Fig.6.14 better, we compare the forecast distribution with the climatological distribution. Fig.6.17 shows

the Relative Entropy (RE) between these distributions for the month of May for 64 years with equally probable (Fig.6.17 top) and equally spaced bins (Fig.6.17 bottom). These approaches for calculating RE were introduced in the section 3.5. The blue line represents the RE between in-sample set and climatology while the red line shows the RE between out-of-sample set and climatology (see Fig.6.17).

Initially (as expected), the forecast is very different from the climatology and RE is the largest. At this stage blending parameter α is high. After 40 years the RE is low (in all 4 cases, consistent with sampling error); this means the target forecasts and climatology are very similar. When α is high then the climatology and the forecast tend to be different (see the equation 3.19). Fig.6.17shows that after around 40 years the target forecast and the climatology are not enough similar, for the predictability to set α to non-zero.

As expected, the RE of in-sample set vs. climatological distribution (blue line) is different from the out-of-sample set vs. climatology (red line). We notice here, that the RE is smaller in case (a) than (b) in the first years of the simulation, because fluctuations are smaller. In case (b) when bins are equally spaced, the variation of bins with fewer members is big, so the RE is large for those bins.

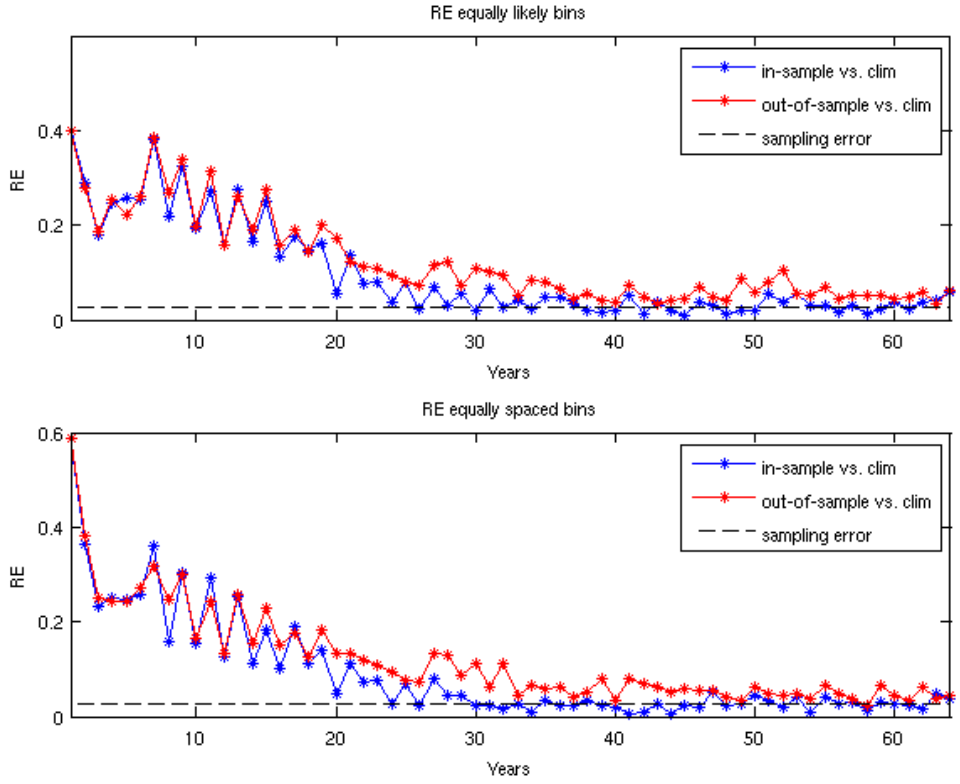


Figure 6.17: Relative entropy between the forecast and the climatology for May as a function of lead time for equally likely (top) and equally spaced (bottom) bins for the in-sample (red) and out-of-sample (blue) sets. After around 40 years, the forecasts and the climatology are very similar and thus the RE is close to zero and around the sampling error (calculated in Tables 6.5 or 6.6). As expected, the RE is smaller in case (a) than (b) in the first years of the simulation, because fluctuations are smaller. In case (b) when bins are equally spaced, variation of bins with fewer members is big, so the RE is big for those bins. RE of in-sample set vs. climatological distribution (blue line) is not the same as out-of-sample set vs. climatology (red line), as both sets are different.

In Fig.6.18 we compare score for in-sample and out-of-sample ensemble forecasts. Markers represent consecutive years in May in all 64 years. All points stay close to diagonal, which means the Ignorance scores calculated on the two different sets are generalising well. This helps to check that in-sample and out-of-sample sets are similar and the derived operational parameters are robust.

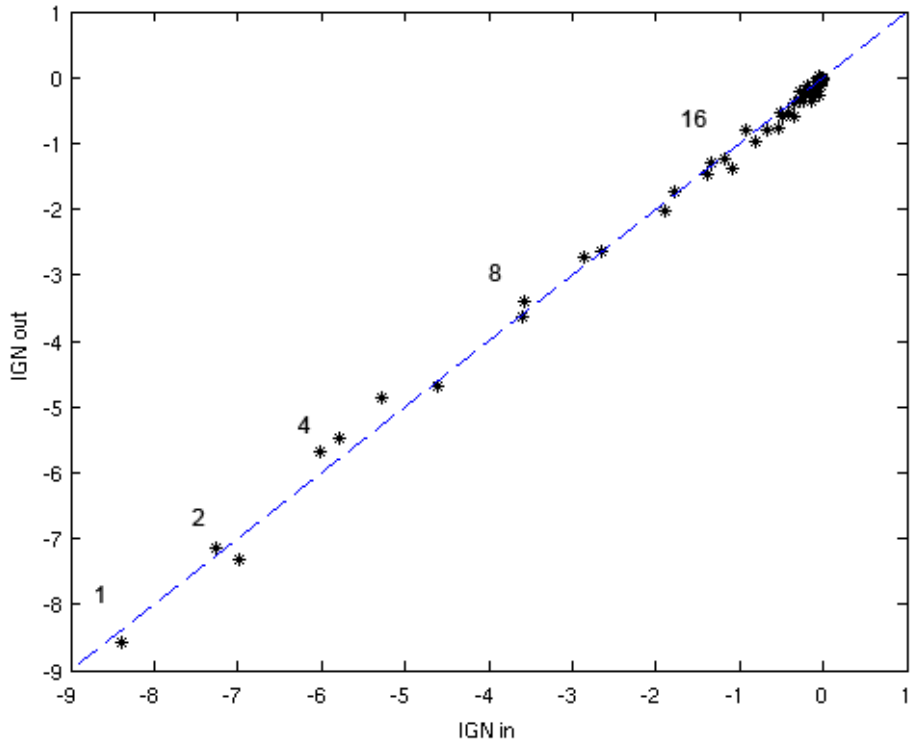


Figure 6.18: In-sample vs. out-of-sample Empirical Ignorance. Markers show consecutive years of May and stay near diagonal. This means in-sample and out-of-sample sets are similar and the derived operational parameters are robust.

6.3.4 Predictability measures applied to the Perfect Model

To show the decay of predictability we use data independent of those used to calculate the parameters. To evaluate the forecast system we compute the Empirical Ignorance, the Model Implied Ignorance and the Information Deficit over the out-of-sample set.

Ignorance of the forecast

Figures shown in this section concern four different months: January, April, July and October, each representing a different season of the year. The Empirical Ignorance

as a function of year for the four chosen months is shown in Fig.6.19. In general, in all cases the Empirical Ignorance approaches 0 after around 30 years whilst, before then, the forecast outperforms the climatology.

The Empirical Ignorance of each individual forecast for October is illustrated in Fig.6.20. For longer lead times, the Empirical Ignorance is around 0 on average. Similarly, to the mathematical model case presented in Fig.5.8, some forecasts have a positive ignorance score. This happens when the target forecast falls outside of the range of the ensemble. This issue is illustrated in the top chart of Fig.6.21. The bottom chart of this figure demonstrates a situation when the target forecast falls outside of the range of climatology. This resulted in a very high climatological ignorance, which for example made the Empirical Ignorance exceptionally low after 50 years (see yellowish line in Fig.6.20).

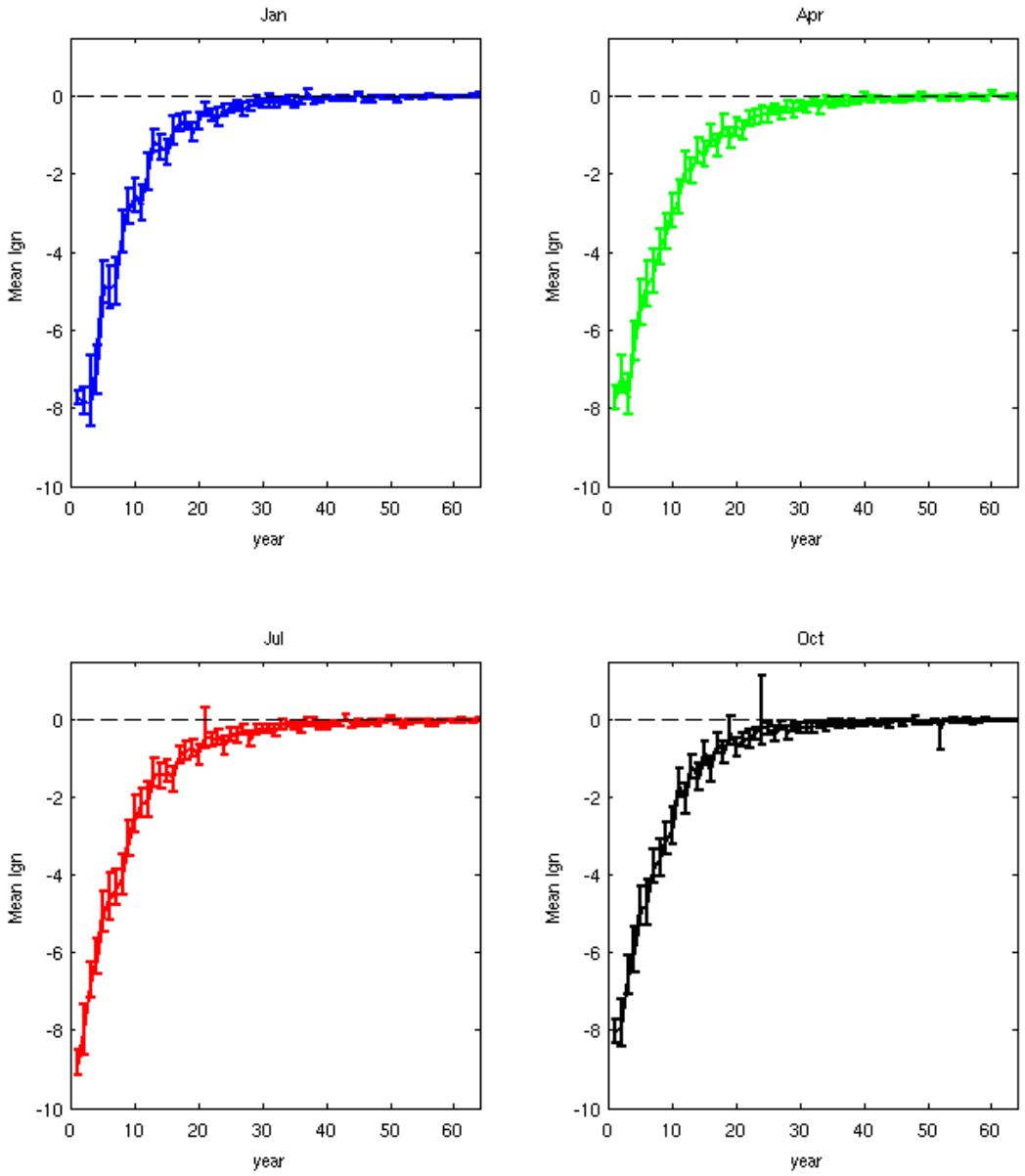


Figure 6.19: Empirical Ignorance as a function of year for Jan, Apr, Jul and Oct in experiment 6.B with 95% resampling intervals. As the lead time increases, the Empirical Ignorance tends to zero and thus the information in the forecast tends to zero. The pattern of the decay is similar to the one on a low-dimensional mathematical model (see Fig.5.16).

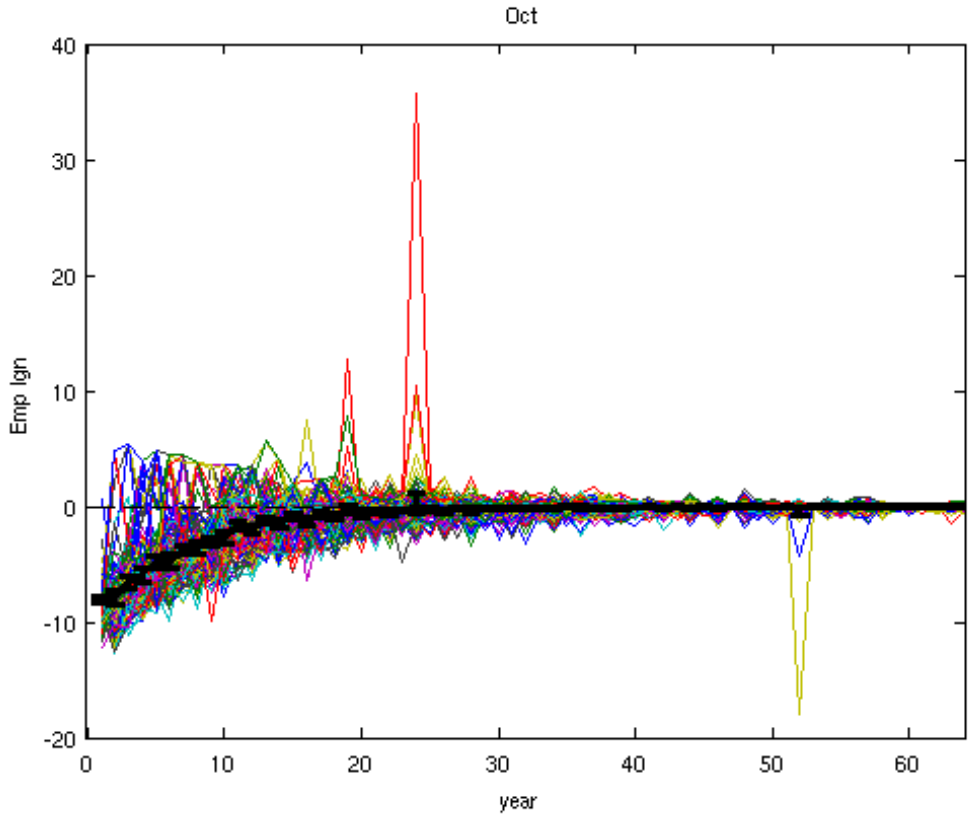


Figure 6.20: Empirical Ignorance for each individual forecast for the month of October in experiment 6.B. Each thin line shows the Ignorance of a different ensemble forecast, whilst the thick black line shows the mean Ignorance over all forecasts. Some forecasts have Ignorance greater than zero for long lead times. This is expected and explained in the top chart of Fig.6.21.

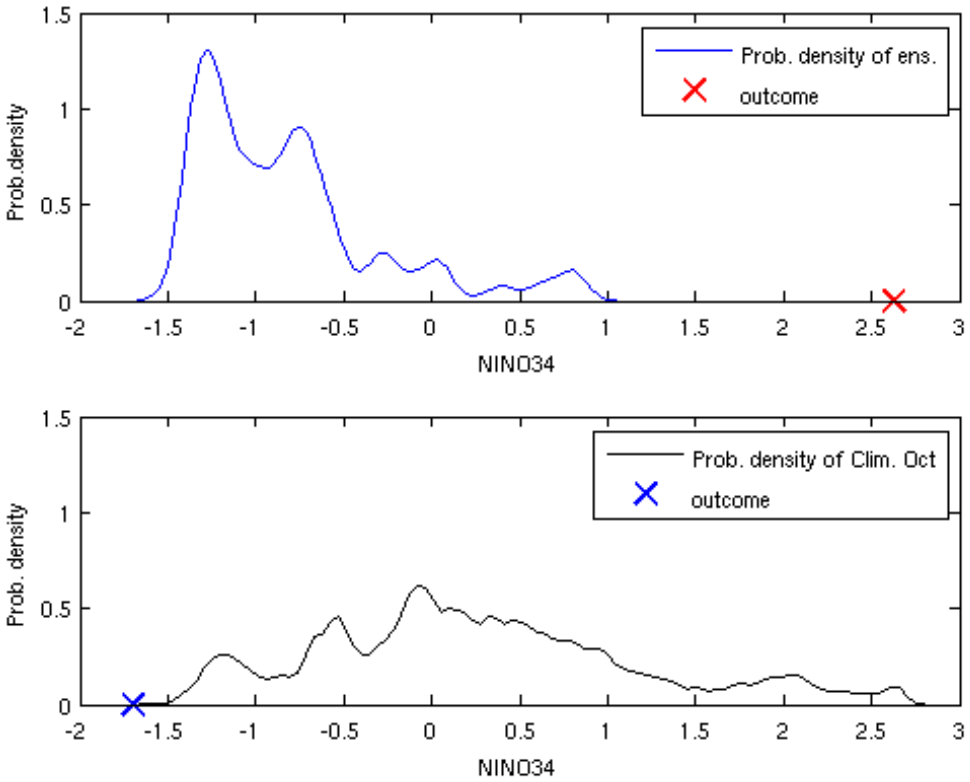


Figure 6.21: Top panel: example of ensemble probability density (blue line) and the outcome (red cross) for a case in which the outcome falls outside of the range of the ensemble. The ensemble at the top, resulted in a positive Ignorance score (see red line at year 25 in Fig.6.20). The bottom chart shows the outcome (blue cross) which falls outside the range of climatology. This resulted in a large Climatological Ignorance, which made the Empirical Ignorance very low (see yellowish line at year 52 in Fig.6.20.)

The Model Implied Ignorance is shown as a function of lead time for forecasts of the four chosen months of the year in Fig.6.22. The Model Implied Ignorance for individual forecasts for the month of October are shown as a function of lead time in Fig.6.23. The Model Implied Ignorance does not increase in a regular manner as in the low-dimensional mathematical model case demonstrated in the previous chapter (see method E in Fig.5.10 or Fig.5.12). At longer lead times, the Model

Implied Ignorance tends to zero since all of our forecasts tend to the climatology.

The value of the Model Implied Ignorance sometimes falls instead of increasing in the consecutive years. For example, when the ensemble forecast falls into states of climatology that are less likely to occur, Climatological Ignorance is large and this will decrease the Model Implied Ignorance. Fig.6.24 demonstrates such an example.

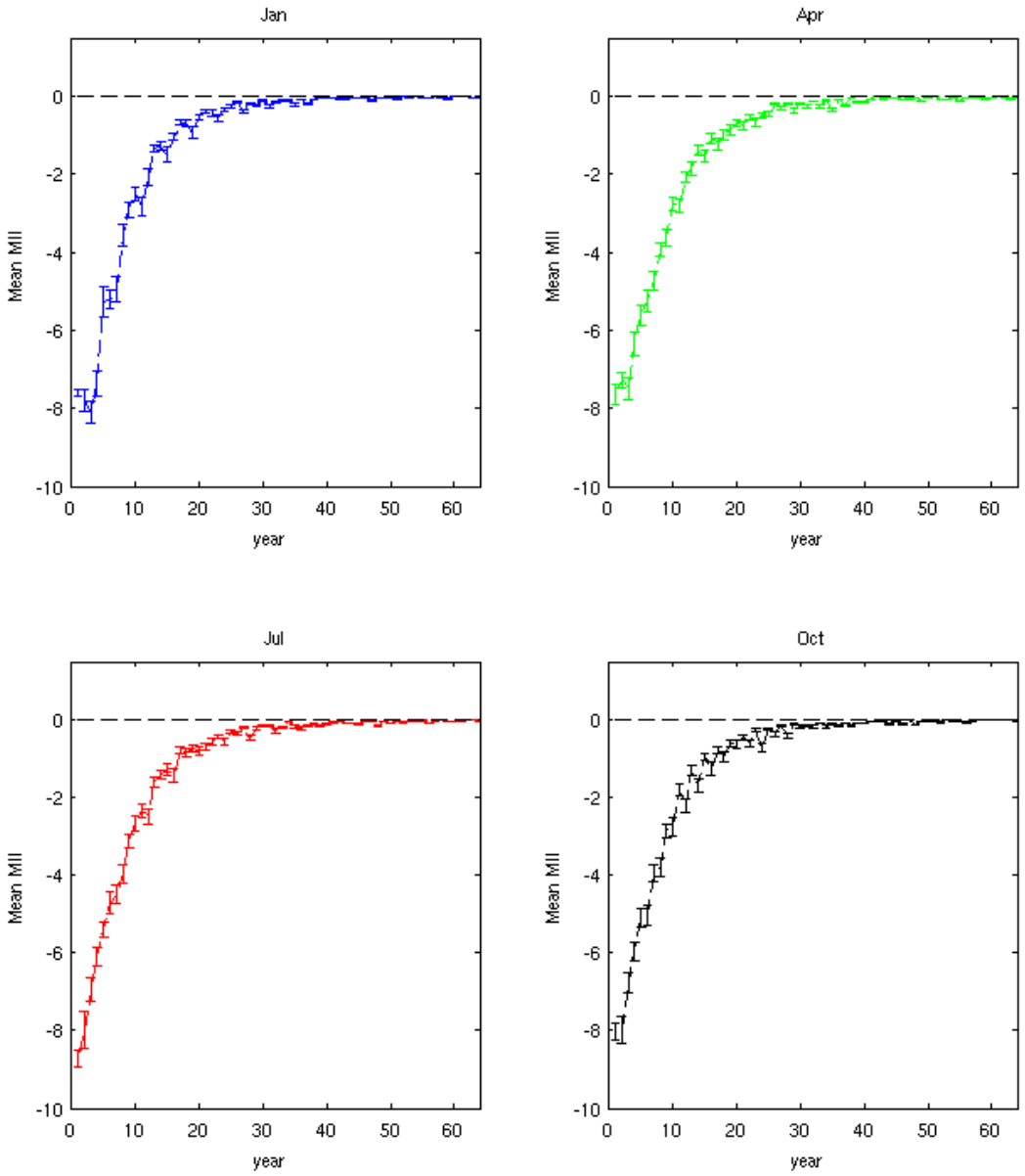


Figure 6.22: Model Implied Ignorance as a function of lead time for four months of the year in experiment 6.B. The Model Implied Ignorance is negative for shorter lead times but tends to zero as our forecasts approach the clmatology. This pattern is similar to the example shown using the simple mathematical model in the previous chapter (Fig.5.10, for example). In contrast to that example however the Model Implied Ignorance does not grow monotonically, but jumps up and down. This issue is explained in Fig.6.24.

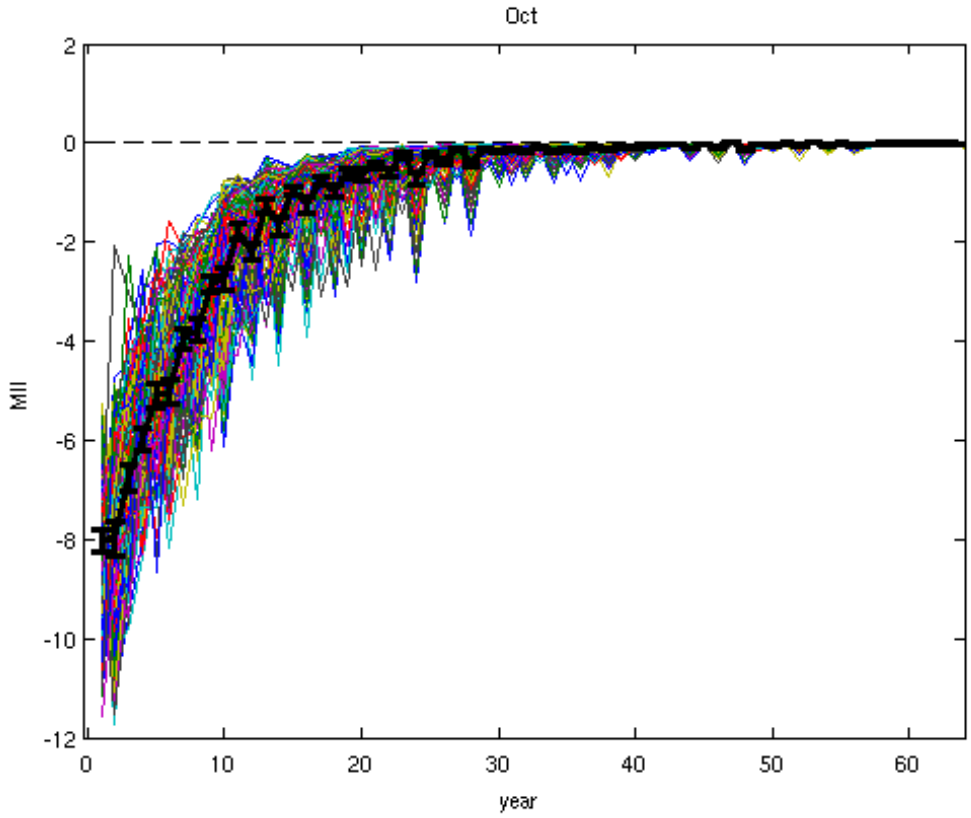


Figure 6.23: Model Implied Ignorance as a function of lead time for the month of October, experiment 6.B. Each thin line shows the Model Implied Ignorance for an individual forecast, whilst the thick line shows the mean over all forecasts. For longer lead times the Model Implied Ignorance converges to zero. In contrast to a simple mathematical model (see Fig.5.12), here the Model Implied Ignorance does not increase monotonically. This issue is illustrated in Fig.6.24. The shape of the teeth is also due to the fact that model has regular oscillations as shown in Fig.6.6.

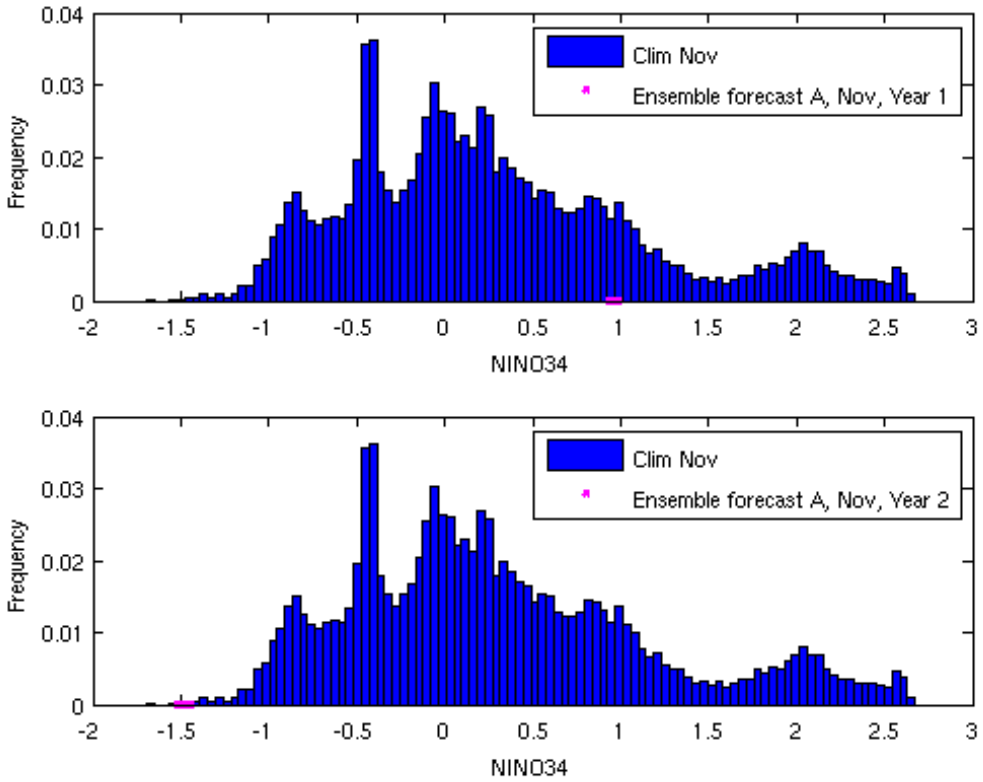


Figure 6.24: Climatology in November (blue bars) and Ensemble forecast A (purple dots) in November at year 1 and at year 2. Ensemble forecast A falls into a range of climatology which puts high probability (top) and low probability (bottom). The ensemble at the bottom resulted in a drop of the Model Implied Ignorance by 6 bits from year 1 to year 2.

Information Deficit

The Information Deficit is the difference between the Empirical Ignorance and the Model Implied Ignorance. The second measure is expected to be better (has lower values) than the first one. In such cases the Information Deficit is positive. When the Empirical Ignorance and the Model Implied Ignorance are the same, the Information Deficit is zero.

The Information Deficit for four different months is demonstrated in Fig.6.25. The

Information Deficit is small for each month. It decreases with increasing lead time and then oscillates around 0 at longer lead times. If we ran the simulation for longer, the Information Deficit would converge to zero. The Information Deficit is never significantly different from zero (95% resampling intervals include zero). This pattern is similar to analogical results based on the simple mathematical model presented in sections 5.2 and 5.3. In the case when the model and the system are the same (green line in Fig.5.15), the Information Deficit is never negative.

Initially however, the results at the beginning of the simulation, showed that the Information Deficit was significantly different from zero, which was not expected. We looked at this issue in a greater detail. First, in order to assess the impact of the climatological part on the Information Deficit, the Empirical Ignorance (equation 3.4) and the Model Implied Ignorance (equation 3.6) were computed without being relative to climatology. In that case the Information Deficit became larger, but it was still negative in a few cases and significantly different from zero. This suggests that error in the climatology is not the main source of error in the prediction.

Next we considered another line of investigation. If a number Y is a random draw from some underlying PDF, the value of the CDF at Y is a random draw from a standard uniform distribution. With this in mind, we can perform a diagnostic test for the forecasts and outcomes in the test set. This is done by plotting a histogram of the value of the CDF of each forecast density at its outcome. If the forecasts are perfect, this histogram should resemble a standard uniform distribution. In fact, although the model is perfect, we do not generally expect the forecast to be perfect since other imperfections such as the ensemble formation scheme will take effect. Nonetheless, we do not expect much deviation of this histogram from a standard uniform distribution since the underlying system density and the forecast density

are expected to be, at least, similar. On looking at the histogram, we noticed that it did not resemble a standard uniform distribution. This helped us to realise that we always defined the outcome as the 65th model run. It thus became clear that the model runs are not exchangeable (i.e. order is important). To fix this, we decided to draw the outcome randomly from the model runs instead. The results are shown in Figure 6.25. This demonstrates how the information deficit can flag up problems either in the forecast methodology or the underlying code.

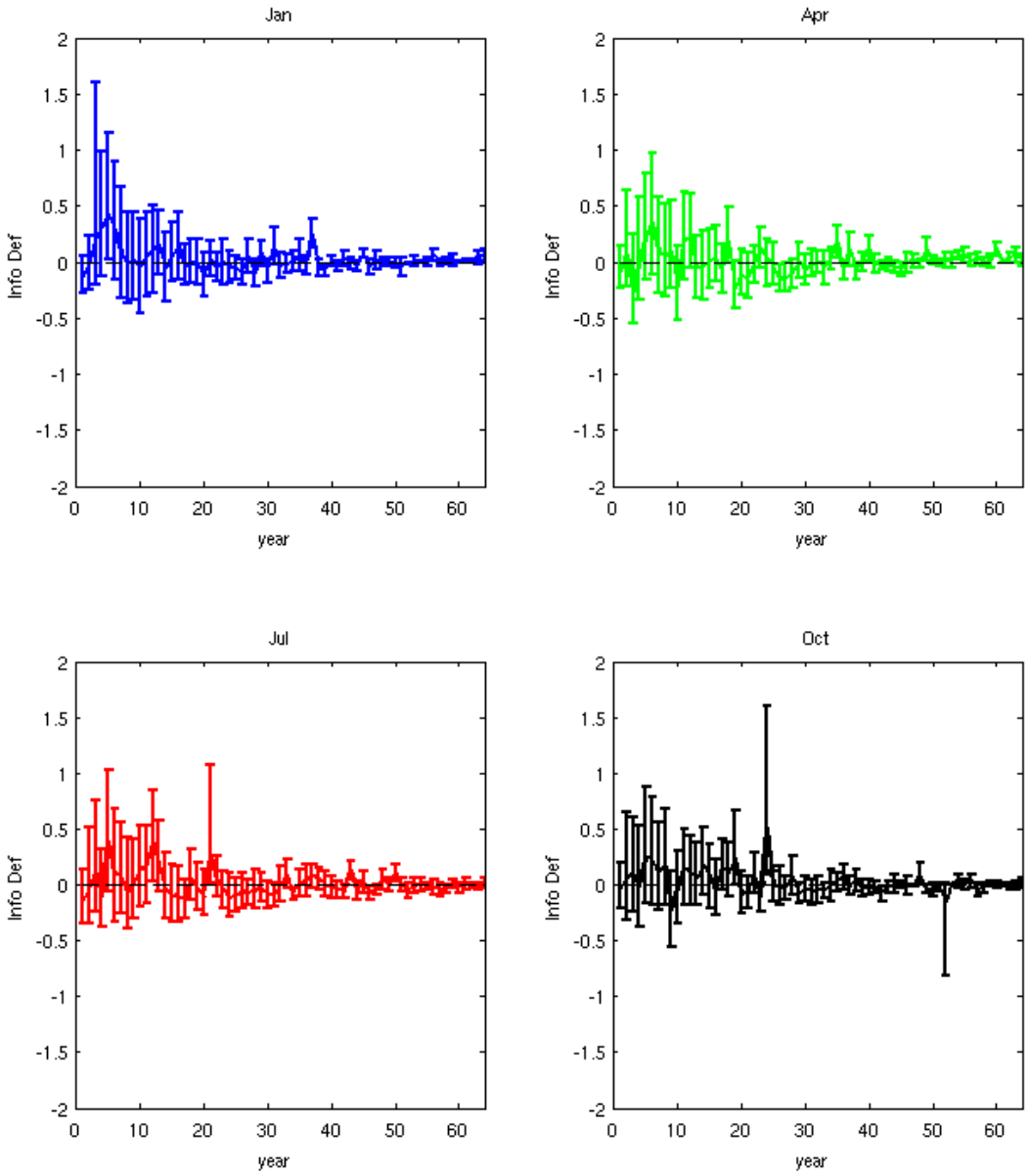


Figure 6.25: Information Deficit in Jan, Apr, Jul and Oct. The Information Deficit is non-negative, as the 95% resampling intervals include 0. In general, the Information Deficit is small, it decreases with years and stays around 0 for longer lead times. If we had data for another 64 years, we would expect that the Information Deficit would converge to 0, as it does in the mathematical model case (see green line in Fig.5.15).

6.4 Predictablity and Forecasting given an Imperfect Model

Here, we consider the Imperfect Model Scenario, presented earlier in 6.2.2. While we talk more about the Imperfect model case in the Appendix C, here we note that it requires the use of a C-Z model with a 5-day time step as the model, and not a 10-day time step (as is used in the Perfect Model Scenario).

We repeat an experiment analogical to 6.B described in section 6.3, but here the ensembles are iterated on the Imperfect Model (5-day model) and each outcome is randomly drawn from the ensembles iterated on the 10-day model. We will refer to this experiment as Experiment 6.E. Details of this experiment are given in Table 6.9.

System	C-Z model (integration time step 10 days)
Model	C-Z model (integration time step 5 days)
n_{ens}	256
Noise level	0.01
m	64
Lead time unit	a month
ts (model)	768 (64 years $\times$ 12 months)
ts (system)	768 (64 years $\times$ 12 months)
n_{icclim}	8
$TRAN_{clim}$	1200 (100 years $\times$ 12 months)
ts (climatology)	22800 (1900 years $\times$ 12 months)
Number of climatologies	12 (for each month)
Size of climatology	15200 (1900 years $\times$ 8 IC)

Table 6.9: Experimental Design 6.E

Fig.6.26 presents the kernel optimal width σ and the weight α for four selected months: Jan, Apr, Jul and Oct. Contrasting Fig.6.26 with Fig.6.14 from a perfect model, we see that α decays faster than in the perfect model case since less information can be extracted from the ensembles.

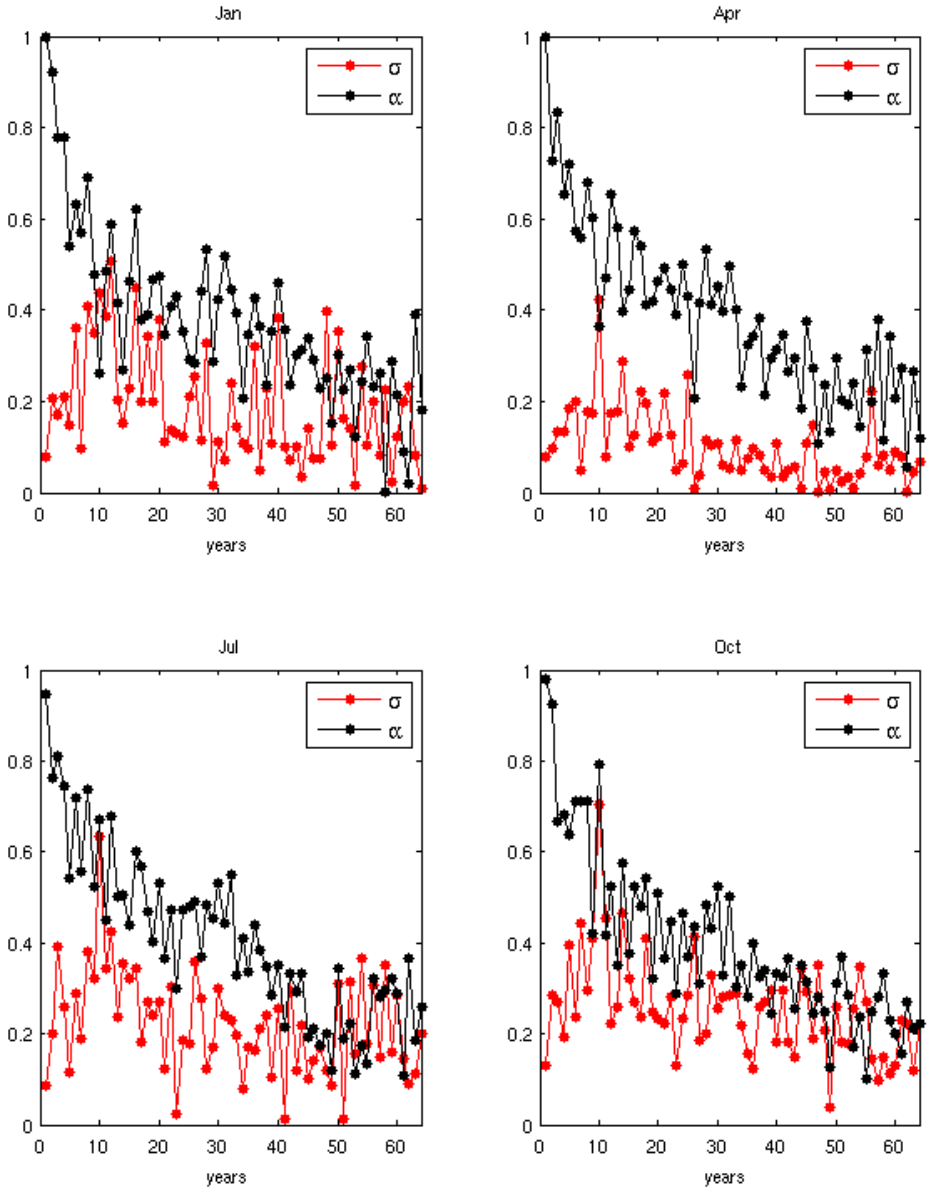


Figure 6.26: Operational parameters for the months of Jan, Apr, Jul and Oct as a function of lead time for the imperfect model case in experiment 6.E. The red line shows kernel width σ and the black line the weight α. Here α decays faster than in the perfect model case (see Fig.6.14).

The graph illustrating the first two predictability measures as a function of year

looks similar to the corresponding PM graphs: (Fig.6.19 and Fig.6.22); however, in the Imperfect Model case the Ignorance decays faster and values at the initial stages of the simulation are larger (between -4 and -3 bits for the Imperfect Model case vs. between -9 and -7 bits for the Perfect Model).

Similar findings were stated in the previous chapter about the decay of predictability in the low-dimensional mathematical model. In section 5.3, it was concluded that in case when the model and the system are the same, it takes longer for the Empirical Ignorance and the Model Implied Ignorance to converge to zero, than when there is a difference between the model and the system. Also, as shown in that section in Fig.5.16, both Empirical Ignorance and Model Implied Ignorance are larger in the Perfect Model case.

Information Deficit as a function of year is illustrated in Fig.6.27. It is non-negative, it is larger at the beginning of the simulation and stays around zero for longer lead times. If the data for another 64 years were used in experiment 6.B, we would expect that the Information Deficit would converge to zero. Here, the Information Deficit is larger than in a Perfect Model case. This finding is similar to the low-dimensional mathematical model example discussed in the previous chapter, namely that the Information Deficit is larger when the imperfection is higher, as is shown in Fig.5.15.

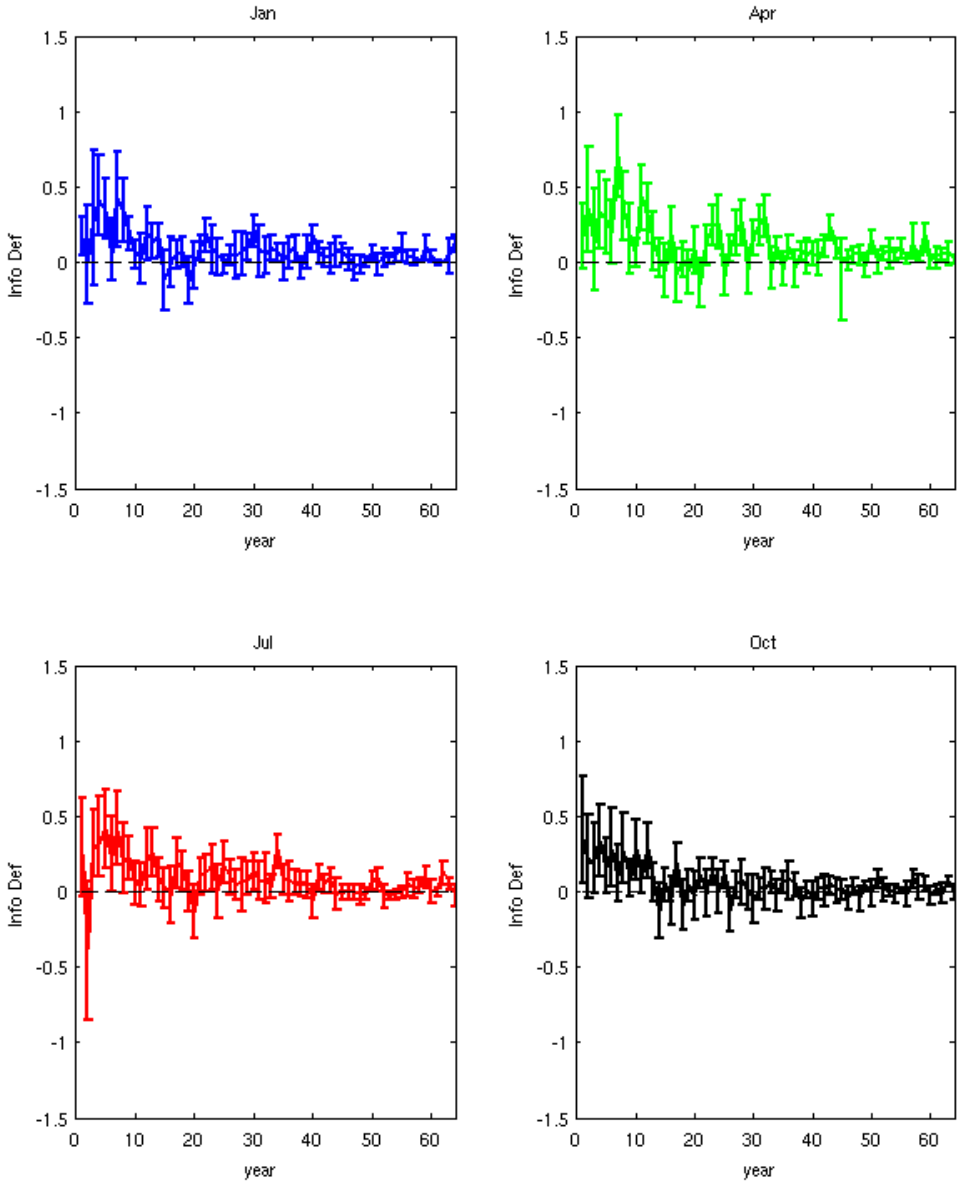


Figure 6.27: Information Deficit for January, April, July and October, Imperfect Model scenario. Here the Information Deficit is larger than in the Perfect Model scenario shown in Fig.6.25, but similarly, it is non-negative, decreases with years and oscillates around 0 for longer lead times. If we ran the simulation for longer, the Information Deficit would converge to 0. Fig.5.15 demonstrates a similar pattern on the Logistic Map, that the Information Deficit is larger for the Imperfect Model case than for the Perfect Model case (for example, compare green line - from perfect model scenario with black line - imperfect model scenario).

6.5 Effect of not having a representative climatology

While calculating the results shown in sections 6.3.4 and 6.4 we encountered some issues, relating to outcomes and some ensemble forecasts being outside the range of the climatology.

Fig.6.28 shows the minimum and the maximum observed value of NINO34 centered in the climatological set along with the minimum/maximum outcome in the out-of-sample set, and the minimum/maximum ensemble member over all ensembles in the out-of-sample set in the perfect and imperfect model cases by month in all years of the simulation. If the climatology was good, then all red stars would be above red diamonds and all blue stars would be below blue diamonds. However, we observe that the climatological range does not systematically cover the range of forecasts or outcomes.

This issue, for example required additional assumptions to be made while calculating Model Implied Ignorance. If the probability density was estimated on the ensemble, we could jump into the situation when $log_2 p_c(x) = 0$ because the climatology is in a smaller range than the ensemble, which results in $\int_{-\infty}^{\infty} p_m(x) log_2(p_c(x))dx = -\infty$. Hence the probability density is evaluated at climatological range. To take into account the ensemble members that are outside the range of the climatology the ensemble forecast is divided by the normalisation constant.

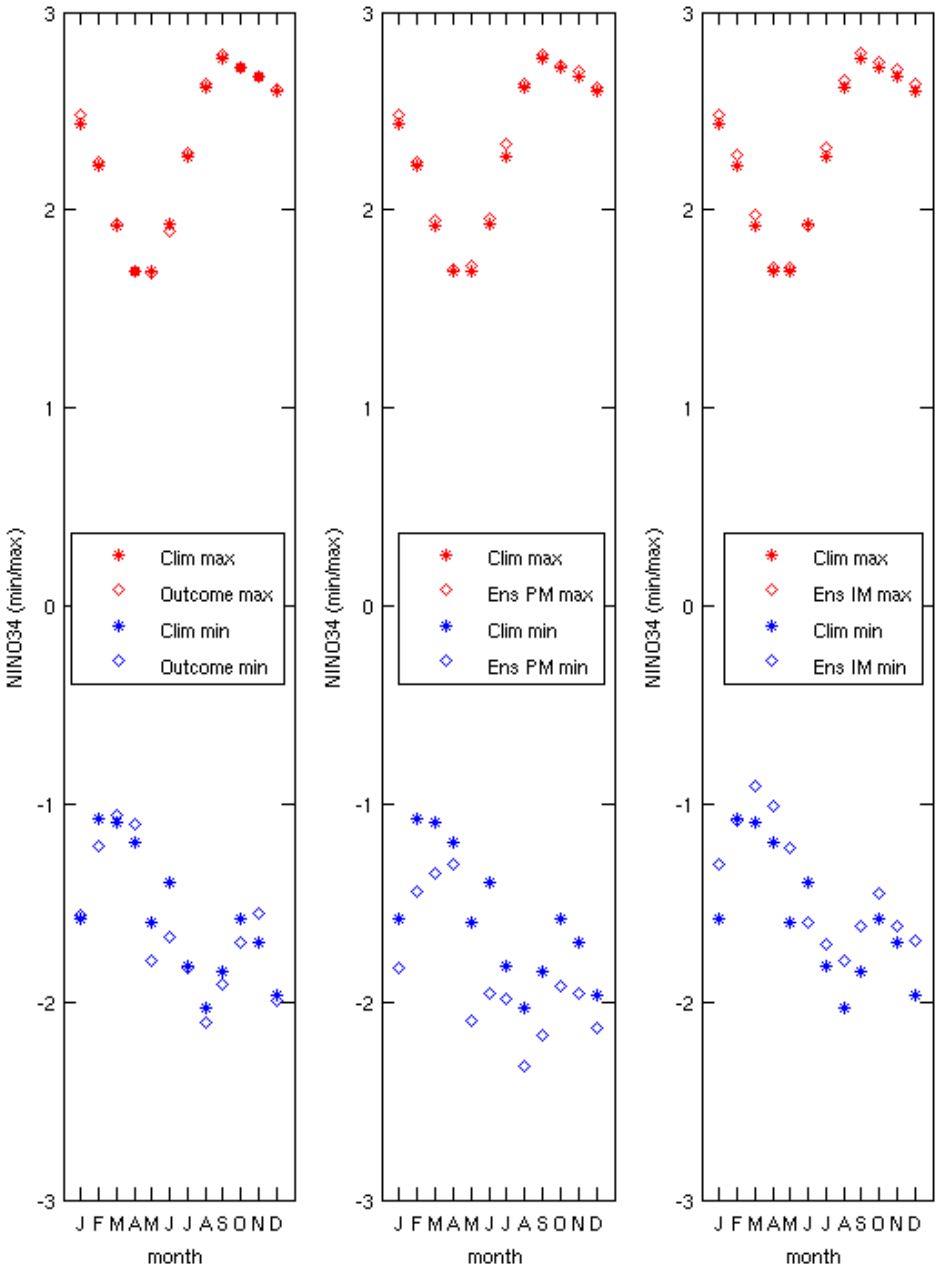


Figure 6.28: Minimum and Maximum value of NINO34 by month for climatological set against (from the left): min/max of outcomes, min/max of ensemble forecasts (from PM) and min/max of ensemble forecasts (from IM). If the climatology was good, we would see in the graphs that red stars are above red diamonds and blue stars are below blue diamonds. However, we observe a different pattern. Climatological range is often not covering the range of outcomes or ensembles. This could result in a large value of Climatological Ignorances and requires additional assumptions while computing the Model Implied Ignorance. Transient states contribute to this difference.

6.6 Conclusions

In section 6.1 we defined El Niño, a phenomenon which has widespread impacts on the world's climate and thus important socio-economic impacts. El Niño influences many areas of human life, hence improved climate predictions could avoid potential losses and damage. This motivated us to study predictability in the context of the Cane-Zebiak model, which is used to forecast El Niño events.

In section 6.2, we focused on the properties of the Cane-Zebiac model which incorporates the basic features of El Niño in the tropical Pacific. The output of the model (NINO34 index trajectories) demonstrates non-linear behaviour and sensitivity to its initial conditions. This suggests that Columbia University's model for El Niño is suitable for application of the non-linear analysis methods, which were developed in the thesis.

In chapter 5, we demonstrated the decay of predictability using low-dimensional mathematical systems. In this chapter, we applied non-linear analysis methods on the much more complex Cane-Zebiak model. In section 6.3, we demonstrated the decay of information in the context of the C-Z model. First we focused on the Perfect Model Scenario (PMS). This is the first time the decay of predictability has been studied in detail in the PMS of the C-Z model. Building the components of the research was far more complex and running the simulation took far more computational time than for the low-dimensional models used in chapter 5.

Section 6.3.2 focused on the climatology of the C-Z model. Given the seasonal nature of the oceans and atmosphere, a different climatology was used for each month. This differs from the low dimensional model used in chapter 5 in which there is no seasonal component. We examined the method of defining the climatology proposed

by scientists from Columbia University. Our results showed that a long time series of transient states is required before non-transient states are reached and suitable to be used to build the climatology. We found that the transient period of 100 years suggested is reasonable.

In section 6.3.3, we studied the kernel dressing parameters used to build probability forecasts from ensembles of the C-Z model. The blending parameter (the weight given to the model-based forecast) was found not to decrease smoothly in time. This was different to the results found for the low-dimensional mathematical models presented in chapter 5, in which the blending parameter drops fairly consistently with the lead time. The reason for this difference is likely due, at least partly, to a higher sampling rate of lead times, which makes sampling error more prominent.

The Information Deficit proved again to be a useful diagnostic tool, uncovering some problems in the forecasting methodology. This time, the issue was caused by the sampling of the outcomes and ensembles from the model output.

In section 6.4, we considered the Imperfect Model scenario in which the model and system were defined with slightly different integration time steps. Whilst, previously, the C-Z model has been studied in scenarios in which parameter error is introduced, this is the first time an experiment has been conducted in a case in which the model is designed to have structural model error. There were challenges with obtaining the imperfect model which are briefly explained in the chapter. As expected, we found that the decay of predictability in the IMS is faster than in the PMS.

Chapter 7

Dynamical downscaling of climate model projections: consequences for predictability

In this chapter we report on an exploratory study of the North American Regional Climate Change Assessment Program (NARCCAP), comparing climate variables on a coarse grid from global climate models (GCM) with the same variables on a finer grid obtained by dynamical downscaling using regional climate models (RCM).

The aim of this work is to consider how the dynamical characteristics of these large physically-based simulation models influence the potential predictability of climate variables. In previous chapters, we have looked at highly simplified low-dimensional mathematical models to illustrate the principles and methods; here we consider a model class which may have tens of thousands of variables and parameters. The initial plan for this work, to apply the methods described in the rest of the thesis,

was cut short by some unexpected results which have generated new insight into the downscaling procedures and a joint paper is in preparation [108] (joint work is noted below).

This chapter is structured as follows. Section 7.1 gives a short overview of the NARCCAP and the aims of climate simulations downscaling, and explains the concept of RCM. Section 7.3 describes the data in NARCCAP. The main results are presented and discussed in sections 7.4. Section 7.6 summarises the conclusions.

7.1 Introduction: Aims of dynamical downscaling of climate simulations

Global Climate Models (GCMs) are widely applied to understand climate change. Their usefulness is, however, limited for officials and the public because they provide the climate projections at the large scale - hundreds of kilometres.

Policy decisions related to climate impacts would benefit from smaller scale projections if such information was reliable. For example, it would be of great use to know whether rainfall in an area is likely to increase or decrease in future; or the expected rate of sea level rise in low-lying areas. This information would be extremely valuable in informing decisions about infrastructure development or suitable policy responses to change, but incorrect information is worse than no information at all, when it leads to misguided decisions with major negative impacts.

GCMs give a global picture of the direction of climate changes, but they typically have 100-200 km resolution (at best), and thus can not take account of local geographical features such as mountain ranges, large lakes, local vegetation patterns (forests, deserts) or the detail of a coastline, all of which have significant influence on

local climate. RCMs are often portrayed as if they act like a physics-based magnifying glass of their driving models for a certain geographical area. RCMs can provide high-resolution simulations, are capable of describing climate feedback mechanisms at the regional scale, and can resolve processes that GCMs are parametrising on their coarser grid. This is a process known as dynamical downscaling, which takes the information provided by the global model in the region of interest and uses this as a boundary condition for a finer-resolution regional model on a denser grid which has more detail of local topography and can resolve dynamical features at a smaller scale.

The regional models are usually driven by a global model in a one-way fashion: no information from the regional model feeds back into the evolution of the global model which drives it. This observation will be key to the discussion which follows, in which we consider some physically significant climate variables and demonstrate that for the models studied here, the regional model diverges from the imposed boundary conditions to such an extent that the one-way driving method is simply unsupportable. If the RCM and GCM were coupled and run together, such that the output of the RCM were to feed back up the scale into the GCM as well as taking the GCM outputs, then we would expect to see significantly different outputs from both.

7.2 The NARCCAP downscaling project

The **North American Regional Climate Change Assessment Program (NARCCAP)** is an initiative of the National Centre for Atmospheric Research (NCAR) in the USA. NARCCAP provides a multi-model high resolution climate simulations over a domain covering most of North America for current (1971-2000) and future

(2040-2070) periods.

The international project NARCCAP provides multi-model output for present and future climate at the model spatial scale of 50 km^2 over the entire North American continent and constitutes the object of our study [70]. The fundamental scientific motivation of NARCCAP is to explore the separate and combined uncertainties in regional climate change simulations that result from the use of different atmosphere-ocean general circulation models (AO-GCMs) to provide boundary conditions for different regional climate models (RCMs) [69]. The aspiration of NARCCAP is to contribute to the research on the value of high-resolution information for preparing for climate change. The program has additional aims: to explore some of the remaining uncertainties in regional climate modelling (e.g., importance of compatibility of physics in nesting and nested models); to generate climate change scenarios for use in impacts research and to create a greater collaboration between the US, Canadian and European climate modelling groups.

Phase I of the NARCCAP project provides information from RCMs run using boundary conditions from the NCEP-NCAR reanalysis for a present-day climate period 1979-2004. Reanalysis is a scientific method, which combines observations and models to produce more detailed information about the recent climate than is available from observations alone. Essentially, reanalysis is a process of generating a climate model run which is consistent with all of the available observations, and is a way of filling in the gaps between the observed points and of generating a gridded data set suitable for use as an RCM boundary condition. More information about the process of reanalysis, and about this reanalysis in particular, can be found online [8] and in the literature [59].

Phase II provides information from RCMs run using boundary conditions from

global climate models (from the CMIP3 [1]) for a present-day climate period 1971-2000 (20th century simulations) as well as from a future time period 2041-2070 based on the SRES A2 emissions scenario for the 21st century. The A2 scenario, which is at the higher end of the SRES emissions of carbon dioxide (CO_2) and other greenhouse gases scenarios, was chosen, as from an impacts and adaptation point of view, it gives more information about the pattern of climatic changes than a low emission scenario would. It results in a doubling of atmospheric CO_2 concentrations (as above, identified with a radiative forcing of approximately $3.7W/m^2$) well before the end of the 21st century [75].

More information about the NARCCAP's experiments can be found on its website [6]. There is a comprehensive description of the program and its experiments, as well as links to download data and some processed output.

7.3 NARCCAP data and definitions of study regions

7.3.1 RCM/GCM pairs used

This research contrasts regional climate variables from a regional climate model with projections for the same region made by the global model driving it. The NARCCAP database is used as a source of our research [6]. Analyses of long term and short term averages of atmospheric variables, such as radiation upwards (rlus), and surface air temperature (tas), are from five of RCM/GCM pairs. These are:

a) Experimental Climate Prediction Center Regional Spectral Model forced by

[1] CMIP3 - Phase 3 of the Coupled Model Intercomparison Project [2]

the Geophysical Fluid Dynamics Laboratory GCM (ECP2-gfdl),

b) Canadian Regional Climate Model forced by the Community Climate System Model (CRCM-ccsm),

c) Canadian Regional Climate Model forced by the Third Generation Coupled Global Climate Model (CRCM-cgcm3),

d) MM5 - PSU/NCAR mesoscale model forced by the Community Climate System Model (MM5I-ccsm),

e) Weather Research Forecasting model forced by the Third Generation Coupled Global Climate Model (WRFG-cgcm3).

Detailed characteristics of the RCMs are summarised on the NARCCAP website [6]. Two of them (RSM and CRCM) use spectral nudging, which provides information from the nesting model not only at the boundaries and initial conditions, but throughout the domain thus allows to constrain a regional model to be more consistent with GCM behaviour. The GCM data come from CMIP3 [2].

All model outputs are saved in NetCDF files, which is a data format used for storing array-oriented scientific data. R has been used to process this data. For this study, averages and totals have been derived from the archived 3-hourly data. A typical NARCCAP output of 5 years for just one variable and only one RCM/GCM combination, contains around 1 GB of raw data to process. For example, for the calculations described in section 7.4.1, 10 GB of data was required.

7.3.2 Definition of study areas

Analysis includes all 0.5 degree grid cells which fall within three areas: USA area, North-West and South-East of the USA. Our region of study, which we refer to as the USA area, encompasses the special domain from 123 °W to 70 °W and 26 °N to 52 °N. The United States, due to its size and range of geographic characteristics, contains many climate zones and thus we also focused on smaller, homogeneous regions. The area defined as North-West (NW) is the area ranging from 37 °N to 47 °N and from 120 °W to 110 °W. It experiences a wide variety of local climates: Oceanic climate in most coastal areas, Alpine climate in high mountains, and Semi-arid climate east of the high mountain ranges. The South-East (SE), which has a warm to hot, humid, subtropical climate, is defined here as the area between 30 °N to 37 °N and from 95 °W to 80 °W. Fig.7.1 shows the study regions.

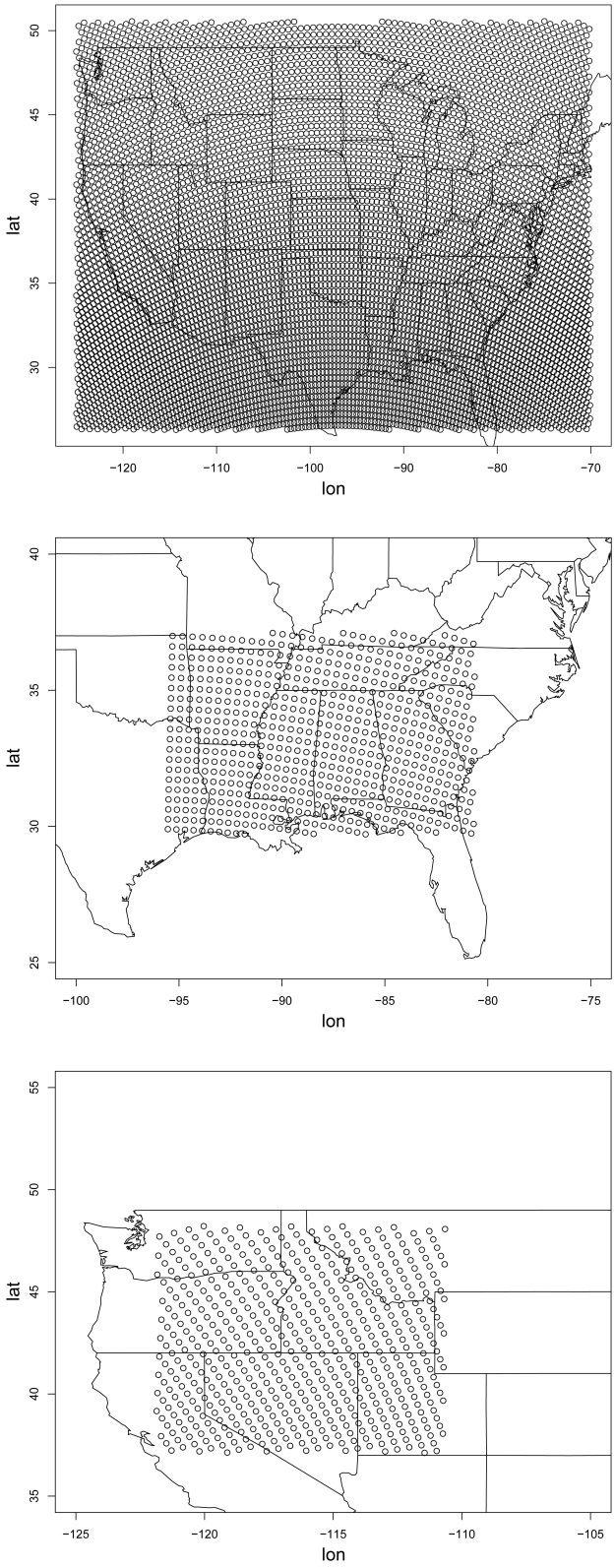


Figure 7.1: Maps showing the study areas as a grid on RCM. In this research we refer to them as (from the top) USA area, South-East and North-West of the USA.

7.4 Comparison of GCM and RCM output for the same areas

This section focuses on investigating the consistency of regional climate projections with the global conditions that stimulated them. To evaluate the significance of the adjustments generated by RCM downscaling, we consider the difference in radiation upward between regional model projections and the corresponding (forcing) global model projections and compare it with the radiative forcing of approximately 3.7 W/m^2 associated with doubled CO_2 (as stated in Box 12.2 of the most recent IPCC report [25] and in TFE.4 of the Technical Summary [106]).

Although local changes greater than this amount are to be expected due to local topography, if the regional or continental average adjustments exceed this magnitude then we would expect significant errors in ongoing forcing conditions due to the neglected feedback effects.

We then consider comparison of precipitation. Again, we expect that the RCM should provide local detail but not to make regional or continental scale average adjustments greater than a small fraction of the total. Otherwise, this would suggest that the large-scale transport and storage of moisture is significantly different.

7.4.1 Comparison of long-term (10-year) upward radiation averages

First, long-term averages (over 1991-2000) of radiation upward are contrasted. Analysis for 10-year averages are carried out for five GCM-driven RCM simulations over the same area of North America (USA area). Results are shown in Fig.7.2.

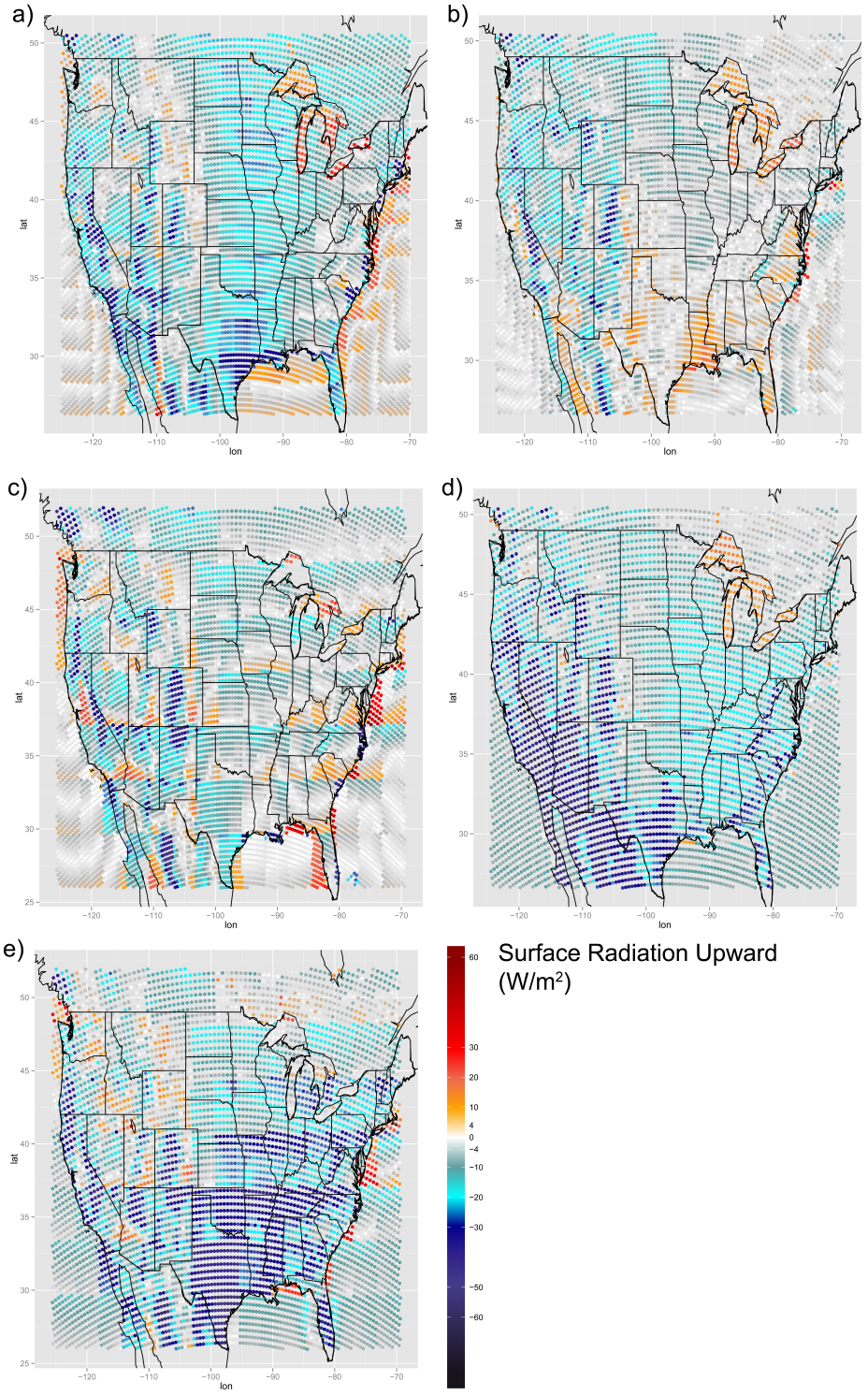


Figure 7.2: Maps illustrate difference in long-term averages (over 1991-2000) of surface radiation upward (variable 'rlus') between RCM output and its driving data by the grid point. The results concern the following NARCCAP RCM-GCM pairings: a) ECP2-gfdl, b) CRCM-ccsm, c) CRCM-cgcm3, d) MM5I-ccsm, e) WRFG-cgcm3. The average difference for the USA area is: -8 W/m^2; -3 W/m^2; -4 W/m^2; -16 W/m^2 and -14 W/m^2 respectively.

We note the following observations:

1. The difference between regional and global model types, is positive for pairs a), b) and d) over the area of lakes on in the North East of the USA. This is expected and indicates that RCMs can be useful tools for capturing physical features of the territory which are not currently resolved by GCMs.

2. In general, all RCMs have lower radiation compared to GCMs (this is expressed by blue coloristics on each map). The average difference in long-term averages of surface radiation upward between RCM and its parent GCM on the RCM grid for the USA area are: a) -8 W/m^2; b) -3 W/m^2; c) -4 W/m^2; d) -16 W/m^2 and e) -14 W/m^2. For comparison, as above, the net radiative forcing due to a doubling of CO_2 is 3.7 W/m^2. This is a worrying observation because it shows that the net energy balance within the region is very different in the RCM from the GCM. A small difference (<0.01 W/m^2) might be acceptable but such a large difference is likely to violate conservation of energy principles and would certainly cause changes in the GCM output if run in a coupled mode.

3. For the pairs that have been examined, those with the spectral nudging feature (cases b) and c)) show a lower difference in long-term averages of surface radiation upward than other GCM/RCM pairs. This does make sense, because the spectral nudging is effectively inserting the GCM data into the middle of the RCM domain, not just at the edges, so it is plausible that this would constrain it to be more like the GCM throughout. The averages are, however, still unacceptably divergent even with this additional constraint.

4. The area of research contains many climate zones and, as expected, results

vary by region. For example, for model combinations a), d) and e), RCM averages differ from those of the parent GCMs over the area of Texas even by as much as 30-40 W/m^2. These large differences would not be unacceptable if they were balanced out by opposite differences elsewhere in the study region, but this is not observed. The RCM is not simply adding detail to an otherwise equivalent regional average generated by the GCM; it is producing a climate which is physically very different.

7.4.2 Comparison of short-term (monthly) upward radiation averages

For each region (USA area, NW, SE), for a single RCM/GCM pair (ECP2-gfdl), monthly averages are computed for Jan, Apr, Jul, Oct in the 33-year time series. Fig 7.3 shows RCM value (x axis) against RCM/GCM ratio (y axis). This particular pair has been chosen, because it does not use spectral nudging and is therefore more representative of the dynamics of the RCM itself.

Looking at monthly averages, the same pattern is observed of systematically lower radiation in the RCM than the parent GCM, since the majority of points lie below the line of equality. As expected, given the heterogeneous climate of the area, diagrams at region level (NW, SE) show bigger variation of results than output at country level. The biggest discrepancies between models are during the summer months, when the surface radiation is the highest. Plotting the full July time series for each year (see Fig 7.4) shows that the downscaled RCM adjustments to the GCM values are around 50-60 W/m^2 in SE, an order of magnitude greater than the effect of doubling global atmospheric CO_2 concentrations. This difference is fairly consistent over time, because the input GCM boundary conditions do not take into account the feedback of the RCM. If run in coupled mode, over time the RCM and GCM outputs might be expected to converge to a new equilibrium taking into account both sets of dynamical processes (unless they were so different that it caused unphysical errors, a runaway loss/accumulation of heat in certain grid cells, or the termination of the model run).

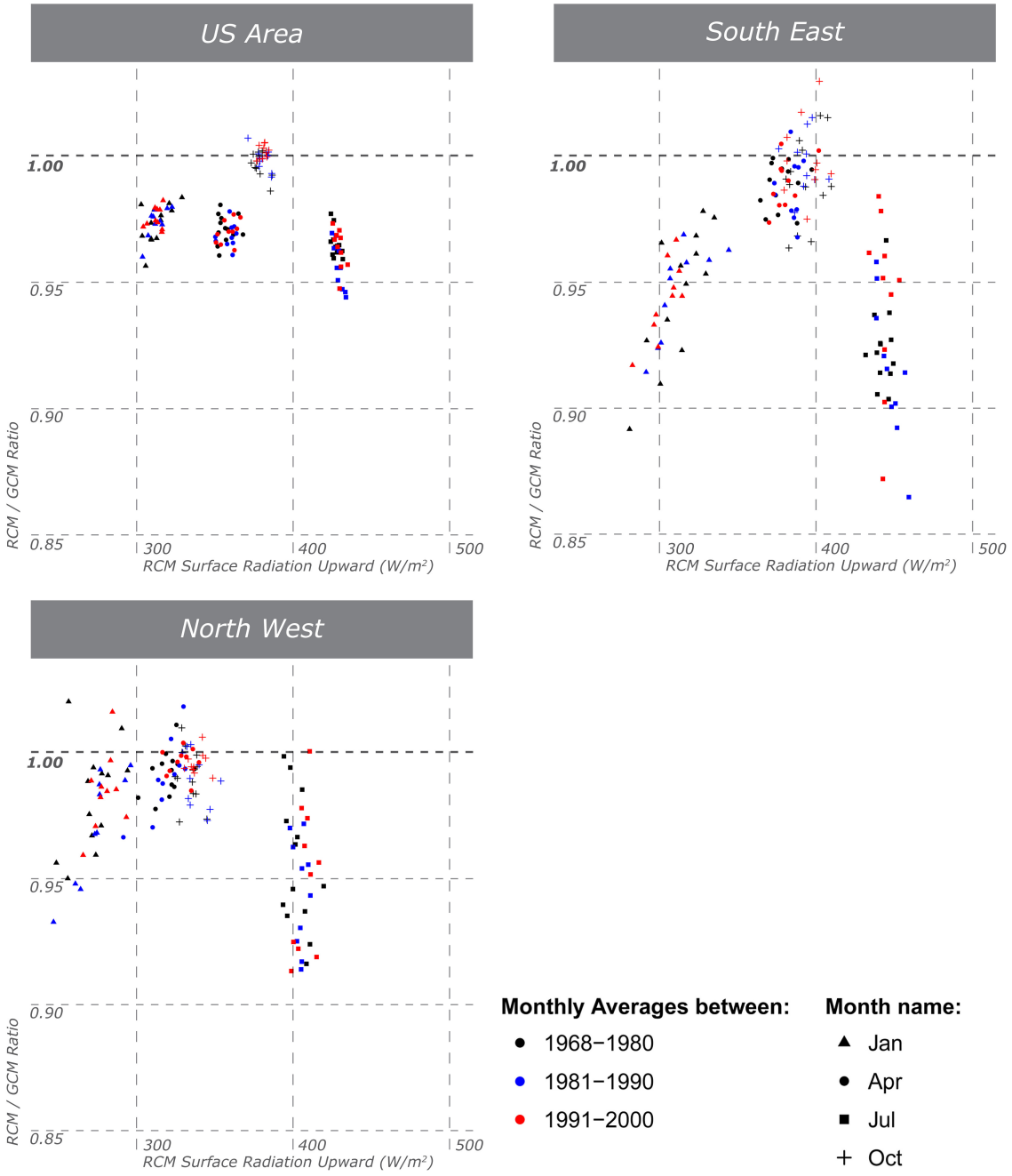


Figure 7.3: Each point in this diagram represents a single monthly average, is coloured by the interval of simulation it falls into (beginning: 1968-1980 -black, middle: 1981-1990 -blue, end: 1991-2000 -red), and its shape depends on the month name. X axis shows RCM surface radiation upward value, Y axis shows RCM/GCM ratio. Charts concern the following regions: US area, North West, South East. Data Source: ECP2-gfdl, NARCCAP. Most points lie below the line of equality because radiation in the RCM is generally lower than in its parent GCM. Results at regional level are, as expected, more varied than at country level, but even over the whole US area, there is a 2-5 % discrepancy in January, April and July on absolute values which are greater than 400 W/m^2, so an absolute discrepancy of over 8 W/m^2. Summer months show the biggest difference in the radiation. For July time series see Fig.7.4.

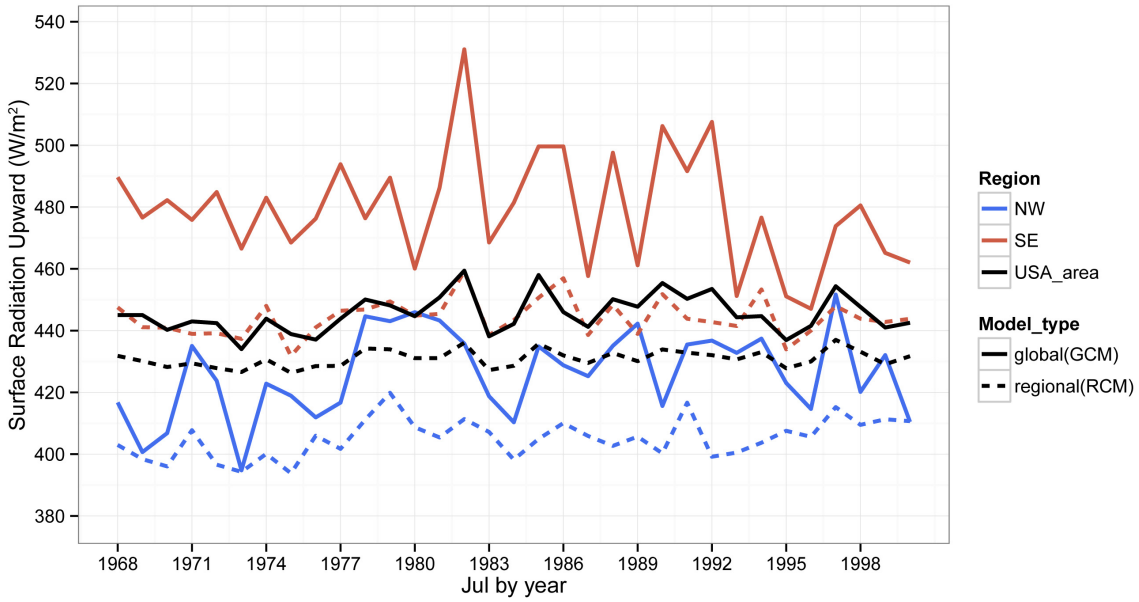


Figure 7.4: July averages of surface radiation upward over 1968-2000 by region and model type. Data Source: ECP2-gfdl, NARCCAP. Note the difference of 50-60 W/m^2 between GCM and RCM in SE. At the country level radiation is on average 10 W/m^2 lower in the RCM than the parent GCM.

7.4.3 Comparison of short- and medium-term average precipitation (joint work with Dr Erica Thompson, LSE)

Note: the work described in this section has been undertaken jointly with Dr Erica Thompson who produced Figure 7.5. We are working on a joint paper which will discuss results shown in this chapter [108].

The surface radiation variable gives us an idea of the energy balance in each model, which will have direct implications for the surface climate. We now consider another variable, precipitation, which is also linked to a physical conservation law (conservation of mass of water) and has strong implications for surface climate and climate impacts. The difference between RCM and GCM output is shown for a single month and a 5-year average in Figure 7.5 (top and middle). The lower panel shows a monthly average precipitation over the whole study area for RCM and GCM, which demonstrates that the RCM generates up to 20% more precipitation than the GCM. This is an indication that the large-scale transport and storage of moisture is significantly different. This in turn would influence the hydrological cycle, local radiative forcing due to water vapour in the air, the growth patterns of vegetation, and the flow of rivers and groundwater. Conservation of water mass must be respected (to some extent), so if significantly more water is precipitating out in a region according to the RCM, then it is either unphysically regenerated downstream by the GCM, or is not available as water in the next grid box. In any case, there is a large and physically important difference here which should be resolved before either model's output is employed quantitatively in science informed decision-making about climate impacts.

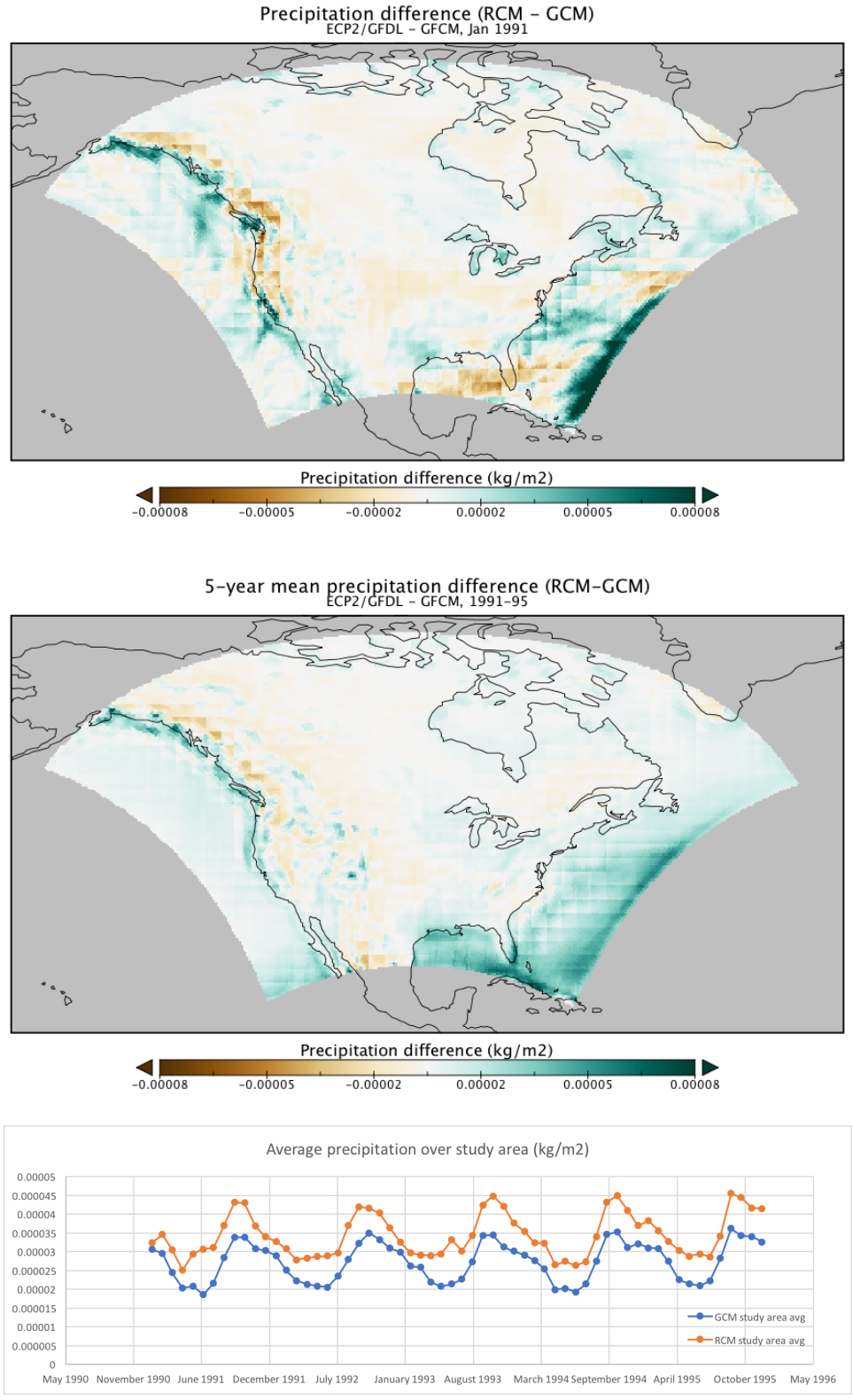


Figure 7.5: Difference between RCM and GCM for precipitation over the NARCCAP study area, for a single month (January 1991), averaged over 5 years (1991-5), and monthly averages over the whole study area. This figure was generated by Dr Erica Thompson of LSE and is gratefully used with her permission [108].

7.5 Does RCM improve GCM output towards the real-world?

Comparison of annual temperature averages at single locations with observed data

The additional computational effort of RCM simulation is usually justified on the basis that it will give a more realistic output, which we take to mean that it will be closer to observed data. Thus the next step is to compare regional model output and its driving data with available observations, in this case for temperature.

The charts below (see Fig 7.6) illustrate annual temperature over the decade of 1991-2000 for four different locations in the USA. The data concerns one RCM-GCM pairing (ECP2-gfdl) from NARCCAP and observed temperature data from NOAA [7].

The centres of the RCM or GCM grids in most cases do not cover the coordinates of the studied location. To find the best match, the nearest point on the grid of the model from longitude and latitude of the place of interest is pinpointed. Fig.7.7 illustrates the studied location on the map (black rectangle) and the corresponding grid centres on the RCM (yellow circle) and GCM (dark green triangle).

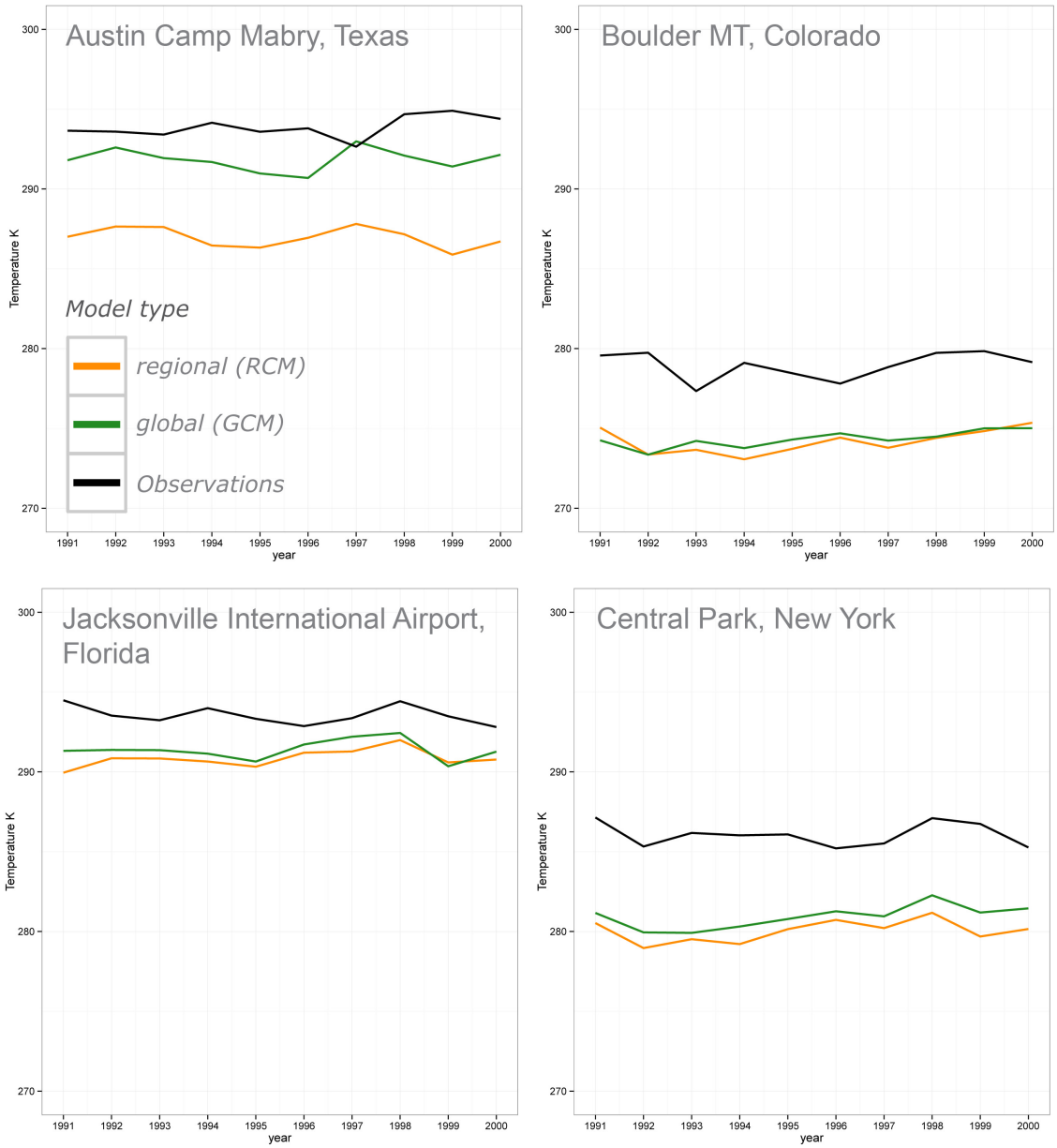


Figure 7.6: Charts illustrate annual averages between 1991-2000 of surface temperature by location. Black line relates to observations, yellow - GCM driven RCM, green- GCM. Locations are: Austin in Texas, Boulder in Colorado, Jacksonville in Florida and New York. Model: ECP2-gfdl, NARCCAP; Observations: NOAA [7]. For all locations, observed temperature was consistently higher than temperature in the models. In addition, the difference in temperature between observations and RCM was bigger than between observations and GCM boundry conditions. Criteria for matching an observation point with the RCM, and GCM grids is explained in Fig.7.7.

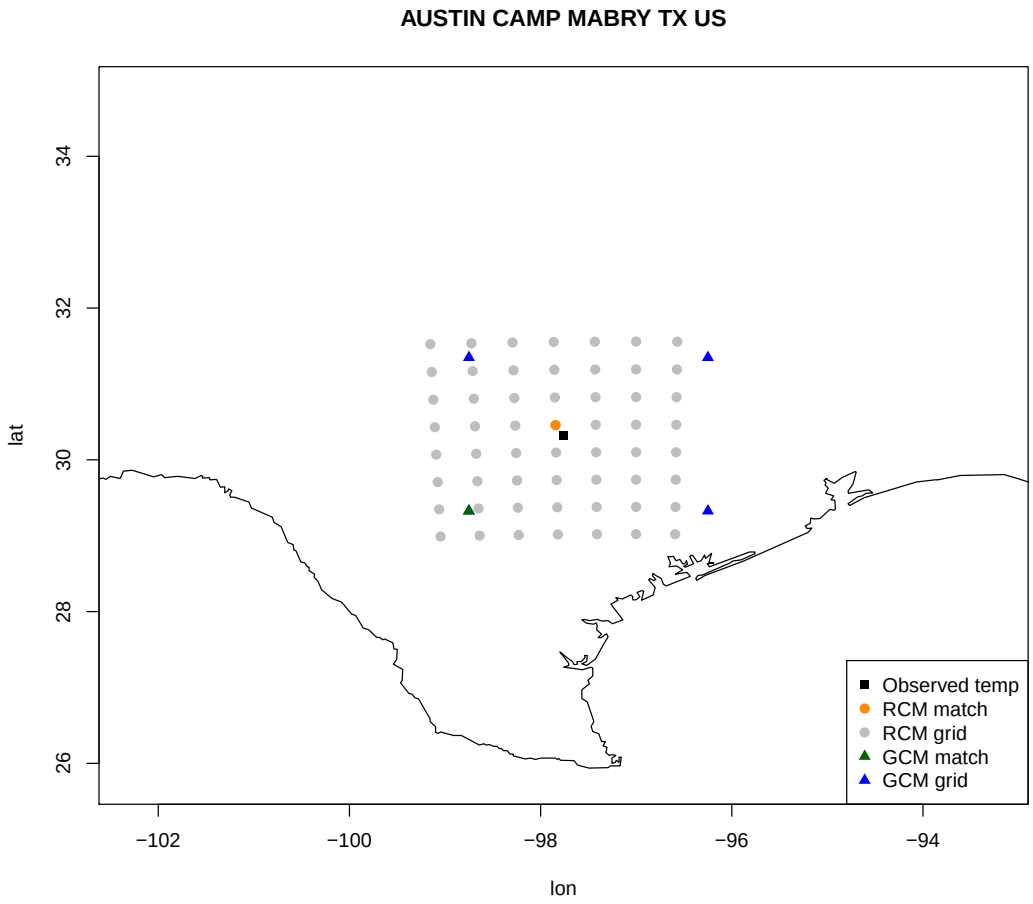


Figure 7.7: Matching an observation point with RCM, and GCM grids. Here, Austin Camp Mabry Texas, US (black rectangle) and the corresponding grid centres on the RCM (yellow circle) and GCM (dark green triangle). RCM grid is illustrated by grey circles, GCM grid by blue triangles. Note that the centre of RCM or GCM grid do not cover the coordinates of the studied location. The best match is the nearest point on the grid of the model from longitude and latitude of the place of interest.

For all locations considered, the observed temperature was higher than temperature in the regional and its driving global models. In fact, for the four locations shown, the RCM data are actually further from observations than were GCM boundary conditions.

7.6 Conclusions

There is a significant divergence between RCM and GCM estimates of both surface radiation and precipitation. The RCM and GCM data deviate to such an extent that the physical relevance of the forcing assumption must be questioned.

An average radiative discrepancy over the whole US above the level of 3.7 W/m^2 is thought to result in significant climatic changes since it is equivalent to the forcing from a doubling of CO_2, a relatively high climate change scenario. It is, therefore, likely to result in significant feedbacks if the RCM and GCM were coupled rather than using the GCM only as a boundary condition. Similarly, a precipitation discrepancy of up to 20% would, over time, result in a very different pattern of vegetation, river flow, agricultural capabilities, etc.

The physical interpretation of output from either model should be approached with caution, given the large sensitivity of the average output to the use of an alternative model. Of course, this applies also at the smaller scale, where the RCM has better representation of topography, but is still subject to an uncertainty in average conditions similar to the range between the GCM and RCM values, which is many times larger than the range generated by doubling CO_2.

Comparison of another atmospheric variable - temperature, between real observations and regional and its driving global models has raised similar concerns to the model-only consideration of radiation and precipitation. It has been shown that RCM data are actually further from observations than were GCM boundary conditions, further undermining the original justification for the downscaling exercise and the associated computational expense.

The original aim of this chapter was to find a time scale on which there is a consistency between regional and global models. Our idea was to look at the lead time at which regional and global models were no longer consistent and examine that period to give an indication of the timescale for which we have reliable predictability. We would then use some of the methods developed in the previous chapters of this thesis. Unfortunately, as described above, the results of initial analysis precluded further study, demonstrating that the surface energy emission between the two model types is inconsistent on very short time scales (in month one).

If that period of consistency was 10 years, we would next have planned to investigate an ensemble interpretation scheme compared with observations and used on the future runs. We would have think about how to do the interpretation of the past, where we actually have data, in order to interpret the ensemble and extract the available information from it.

Under climate change scenarios, we depend on physical insight; we do not want to find statistical ways of using the past and future unless they also have internal physical consistency. Our study of NARCCAP shows we do not have this consistency. The results of this chapter imply that the information we are looking for is not to be found in this set of models. We have gained important information and insight about the use of RCMs to downscale climate information from GCMs and have shown that in the case of NARCCAP, the models are not yet good enough to show any useful predictability horizon. More recent examples of downscaling are available and the joint work with Dr Thompson will be continued, to assess whether these examples show better consistency. In this case it will be possible to use other methods developed in the chapters of this thesis.

The wider aim of this research was to assess the quality and reliability of climate

simulations and the effectiveness of various downscaling methods, in order to better understand the relationship between climate model output, downscaled output, and the climate system itself. We hoped that improving our understanding of these relationships would (i) allow a more informed measure of the fidelity of climate model simulations, (ii) assist in the development of more accurate models, and (iii) aid scientific support for decision-making and policy. We now believe that there are significant physical constraints on the consistency of the climate simulation and downscaling methods employed by the NARCCAP project. We conclude that these simulations, in their present form, can tell very little about future US climate and are not a reliable source of information for policy decisions. More recent progress in downscaling methods may address or resolve these consistency issues and this will be investigated as ongoing future work. The results described above are to be published as part of a joint paper [108].

Chapter 8

Conclusions

In this chapter, we review the main conclusions of the thesis and suggest further work for this research.

The central theme of this thesis is quantifying predictability of nonlinear systems through probabilistic forecasting. We use a measure of forecast skill called Ignorance to access the performance of a variety of different models. Two distinct sources that limit predictability of nonlinear systems, that is structural model error and uncertainty in the initial condition, are identified and considered as separate issues throughout the thesis. Theoretical concepts developed in my research were applied both to a large scale model designed to forecast El Niño and to smaller scale low-dimensional mathematical models such as the Henon Map, the Ikeda Map, the Logistic Map and the Quartic Map.

Some of the most important novel results in this thesis are listed below:

1. Distinguishing model inadequacy and sensitivity to initial condition uncer-

tainty in nonlinear systems. This was argued to be capable of leading to more effective resource allocation in dealing with these two limits of predictability.

2. Demonstration of varying predictability of a chaotic system according to the location on the attractor at the launch time. Identifying when a skillful forecast can be expected a priori.

3. Recognition of shortcomings in the existing forecasting literature on quantifying predictability. Proposal and demonstration of a new improved mathematically consistent approach.

4. Exploration of an existing measure called the Information Deficit and demonstration that it can be used as a diagnostic tool to identify potential improvements to a forecasting system.

5. Studying the decay of predictability of a real-world operational model (Cane-Zebiak model [21]) in both the perfect and imperfect model scenarios.

6. Showing that Regional Climate Models (RCM) are inconsistent with Global Conditions that simulated them in NARCCAP [6] and thus that the output from RCMs should be treated with a high degree of caution in decision making.

The following are some further directions of work that arise as a result of the research, which could be pursued in the future:

1. Explore alternative methods to demonstrate differences between two attractors such as minimum spanning trees (MST) [94] and shadowing times [40].

2. Apply the Information Deficit to forecasts of real world observations to quantify the extent to which the skill of the forecast compares to that expected if

the result were to be drawn from the forecast distribution. Aim to establish guidance so that this could become common good practice.

3. Assess the performance of the C-Z model with real-world observations. Develop a methodology to improve model based forecasts and better interpretation of forecasts for decision support.

4. Develop a consistency test to identify the time scale on which there is consistency between regional and global models to improve experimental designs for potential future downscaling experiments.

Nomenclature

α	blending parameter
$\mathbb{S}$	state space
σ	bandwidth of a kernel
$\tilde{F}$	system dynamics
$\tilde{x}$	system state
ε	distance between two points
d	euclidean distance
F	model dynamics
IC	Initial Condition
IGN	Empirical Ignorance
IGN_{ClimD}	Climatological Ignorance method D
IGN_{ClimE}	Climatological Ignorance method E
IGN_{MI}	Model Implied Ignorance

Inf_{Def} Information Deficit

K kernel function

m number of ensemble members

n number of forecasts

n_{ens} number of ensembles

n_{icclim} number of initial conditions climatology

n_{ic} number of initial conditions

n_{sets} number of sets

$p(.)$ forecast density

p_c density of the climatological distribution

p_m model density

q system density

RE relative entropy

S scoring rule

S_d standard deviation

$TRAN$ duration of transient

ts time steps

Y forecast outcome

x model state

Appendix A

System-Model pairs used in Chapter 4

A.1 Henon Map

The Henon Map [49] is a discrete-time dynamical system, originally introduced as a simplified model of the Lorenz 63 system. The map takes a point x_i, y_i in the plane and maps it to a new point as described in equations A.1 and A.2.

$$x_{i+1} = y_i + 1 - ax_i^2 \tag{A.1}$$

$$y_{i+1} = bx_i \tag{A.2}$$

The map depends on two parameters: a and b, which have the classical values $a = 1.4$ and $b = 0.3$. For these values, the Henon Map show chaotic behaviour. The

attractor of the Henon Map is shown in Fig.A.1 (a).

A.2 Senon Map

The Senon Map is a modified version of the Henon Map, which includes a seasonal cycle described by equations A.3 - A.5 in 'What might we learn from climate forecasts' by L.A. Smith [93]. It is a simple non-linear two-dimensional chaotic map.

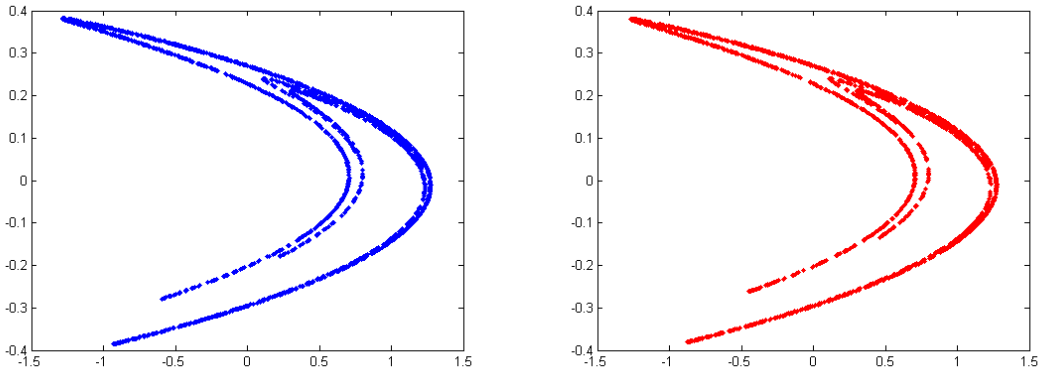
$$\tilde{x}_{i+1} = 1 - a\left(c\sin\left(\frac{\tilde{x}_i}{c}\right)^2\right) + \tilde{y}_i + e\tilde{v}_i \tag{A.3}$$

$$\tilde{y}_{i+1} = b_i c\sin\left(\frac{\tilde{x}_i}{c}\right) \tag{A.4}$$

$$b_i = \bar{b}\left(1 + dcos\left(\frac{2\pi i}{12}\right)\right) \tag{A.5}$$

where a, c, d and e are parameters and $\tilde{v}$ is a random draw from a standard normal distribution.

We consider a purely deterministic system by setting e=0 and set the other parameter values to a=1.4, b=0.3, c=10, and d=0. Note since we set d=0 there is no seasonal factor in equation A.5. The attractor of the Senon Map with these parameter values is shown in Fig.A.1 (b).



(a) Henon Map (b) Senon Map

Figure A.1: Attractor of (a) classical Henon Map and (b) Senon Map [93] with parameters as described above. These attractors look very similar, but yield different results as described in experiment 4.A. on distinguishing model inadequacy and chaos.

A.3 Ikeda Map

The Ikeda Map is a discrete-time dynamical system, originally introduced by Ikeda [53] as a model of laser pulses in an optical cavity. It is described by the equations:

$$x_{i+1} = \gamma + u(x_i \cos\phi - y_i \sin\phi) \tag{A.6}$$

$$y_{i+1} = u(x_i \sin\phi + y_i \cos\phi) \tag{A.7}$$

where $\phi = \beta - \alpha/(1 + x_i^2 + y_i^2)$. For parameters $\alpha = 6$, $\beta = 0.4$, $\gamma = 1$ and $u = 0.83$ the Ikeda Map is believed to be chaotic. Fig.A.2(a) shows the attractor of the Ikeda Map with these parameter values.

A.4 Truncated Ikeda Map

The truncated Ikeda Map can be used to obtain an imperfect model of the Ikeda Map. It was defined by Judd and Smith [96] and it is formed by replacing trigonometric functions of Ikeda Map equations with truncated power series. In this thesis the following truncations are used in experiment 4.B

$$\cos\phi = \cos(\omega + \pi) \rightarrow -\omega + \omega^3/6 - \omega^5/120 \tag{A.8}$$

$$\sin\phi = \sin(\omega + \pi) \rightarrow -1 + \omega^2/2 - \omega^4/24 \tag{A.9}$$

The attractor of the Truncated Ikeda Map is shown in Fig.A.2 (bottom).

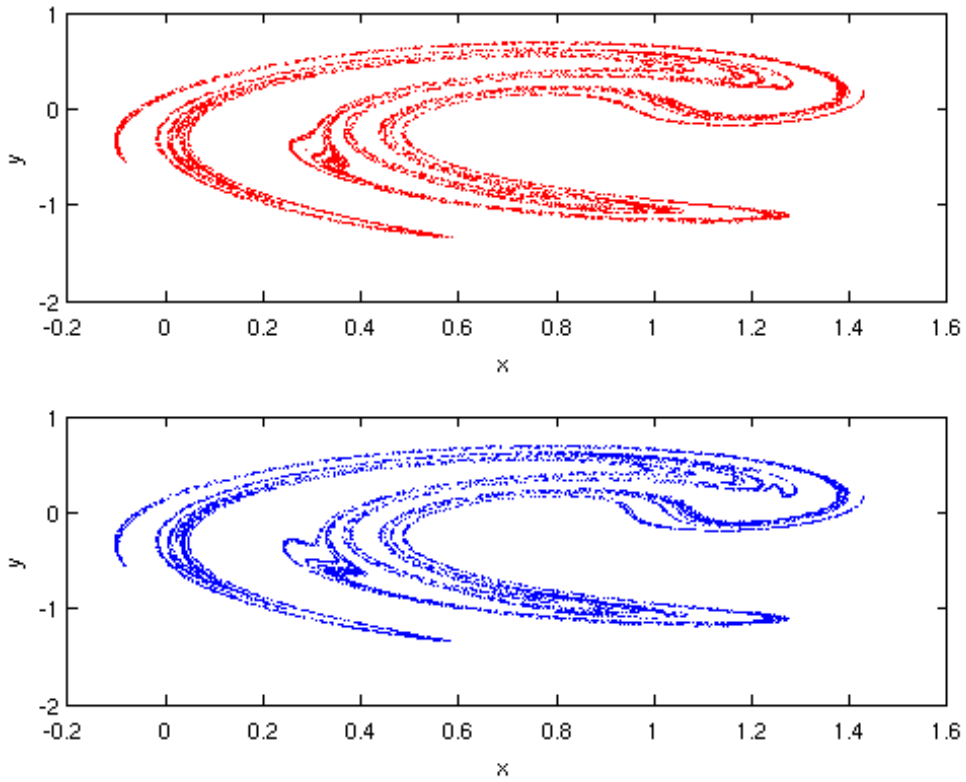


Figure A.2: Attractor of Ikeda Map (top chart) and Truncated Ikeda Map (bottom chart) with parameters as described above.

Appendix B

Climatology of the Quartic Map

Different ways of deriving climatology were considered in chapter 5.

1. Sample Climatology

   In the experiments 5.D and 5.E climatology was derived as 2048 U(0,1) Initial Conditions iterate them for 128 steps as a transient (using system equation 5.2 with different options of μ), continue for 2048 steps recording the distribution of locations of x. This yields 2^{22} number of points evolved on the system. The probability density for each bin of size of truncation between 0 and 1 is calculated. An example of the sample climatology for μ=0 and μ=0.5 is shown in Fig.B.1.

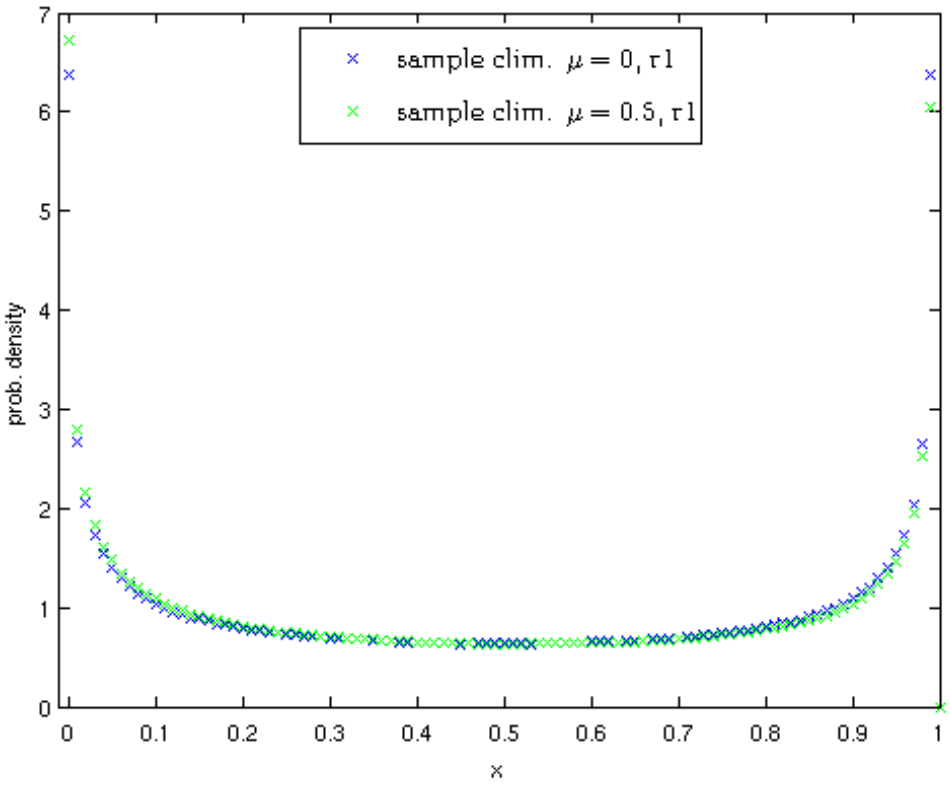


Figure B.1: Sample climatology when μ=0 and μ=0.5. There is a noticeable difference between probability density for both climatologies at the edges of x.

The system with r=4 and μ=0 is equivalent to the Logistic Map with r=4. In this case if an initial state is in (0,1), the attractor is also the interval (0,1) and the probability measure corresponds to the beta distribution with parameters a=0.5 and b=0.5 [11]. This is an analytical form of what happens when the distribution of points is evolved forward.

Fig.B.2 is a comparison of the analytical method and the sample climatology when α is 0. The probability density looks similar in both cases, which justifies the use of the sample climatology in experiment 5.D, instead of beta function with a=0.5 and b=0.5 for the system with μ=0.

This also demonstrates that when the climatology is calculated empirically it does not represent true distribution of past states ('true climatology'). This means that when using climatology as benchmark the forecast skill can be overestimated. However, given that our ensemble size is large, this is unlikely to make much difference to the results.

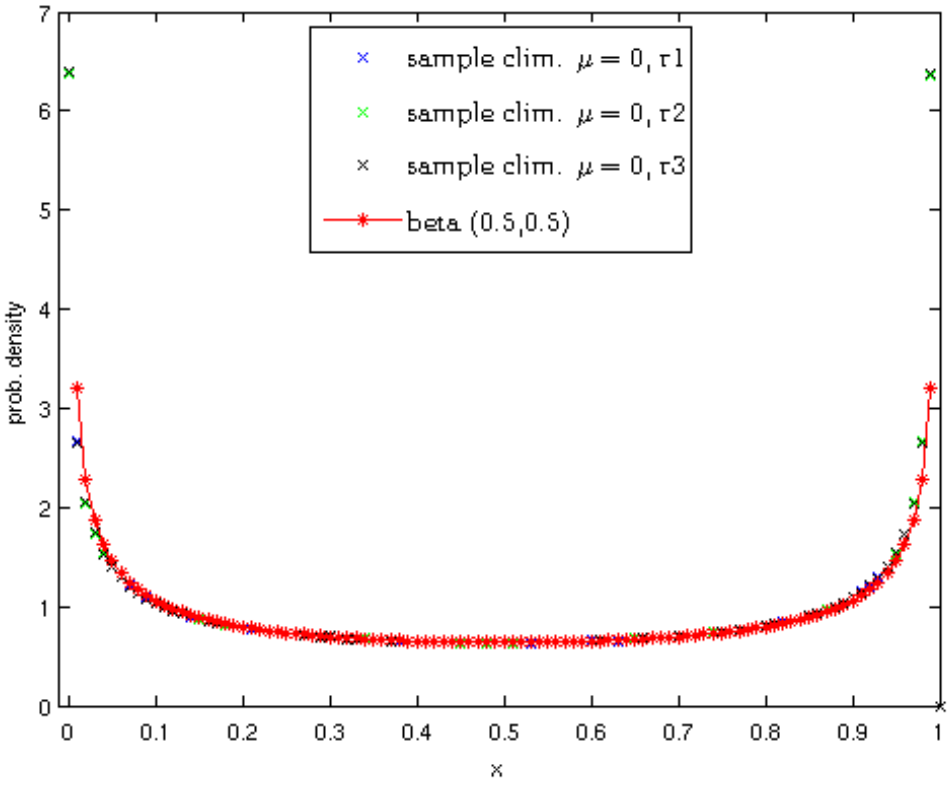


Figure B.2: Sample climatology when μ=0 vs. beta distribution with a=0.5 and b=0.5. Colour of crosses show different runs of the sample climatology. The difference between runs is hard to distinguish. Stars show probability density by beta function for x=0.01:0.01:0.99. Sample climatology is in line with beta function (red curve). As expected, crosses which show probability density at the edges of each bin, are on the beta function curve, in the middle between red stars.

2. Climatology derived by evolving initial conditions on the system for a long time

Another way to derive climatology is: derive 32 U(0,1) Initial Conditions, iterate them for 128 steps (using system equation 5.2) to ensure the states are not transient, then continue for 1024 steps recording the distribution of locations of x. This yields 2^{15} number of points evolved on the system. Then we could estimate the probability density of the climatology using kernel density estimation as shown in Fig.B.3.

We found that the sample climatology is better for further computations. For example, to derive fitting parameters for kernel dressing using sample climatology takes half as much time, than using climatology derived on the system as described above.

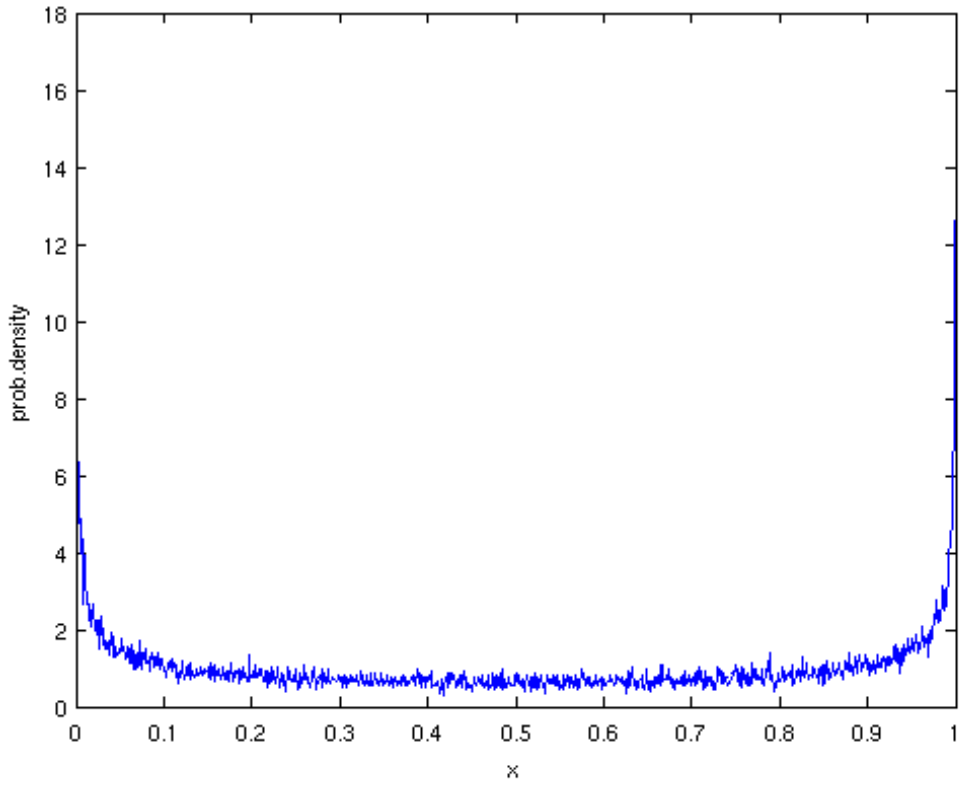


Figure B.3: Probability density estimate of climatology derived by evolving set of points for a long time on the system. This was estimated using kernel density estimation.

Appendix C

Details of Imperfect Model of C-Z model

To obtain an Imperfect Model, the C-Z model is run with an integration time step of 5 days instead of the default 10 days. Scientists from Columbia University, who work on this model did not recommend changing the default setting, as potentially some parts of the code are hard-coded under the assumption that the integration time step is 10 days. We ran a 7-day and 20-day model, but the output showed unreasonable results, such as only zero values in the output of the simulation or extremely high values of the NINO34 index time series. It was also suggested that an alternative way of obtaining an Imperfect Model is to modify one of the parameters in the model, for example the parameter affecting surface heat flux (a linear damping on sea surface temperature anomaly). This, however, would not be an Imperfect Model, since the model structure would remain unchanged. This would, in fact, be parameter error.

C.1 Comparison with the 10-day model

We have already contrasted how the 10-day model and the 5-day model evolve in Fig.6.7. Fig.C.1 illustrates an example of ensemble forecasts for 10-day and 5-day models. For the 10-day model (top chart), the ensemble members start to spread quicker than for the 5-day model (bottom chart). Also, the range of NINO34 values is narrower in the bottom chart than in the top chart.

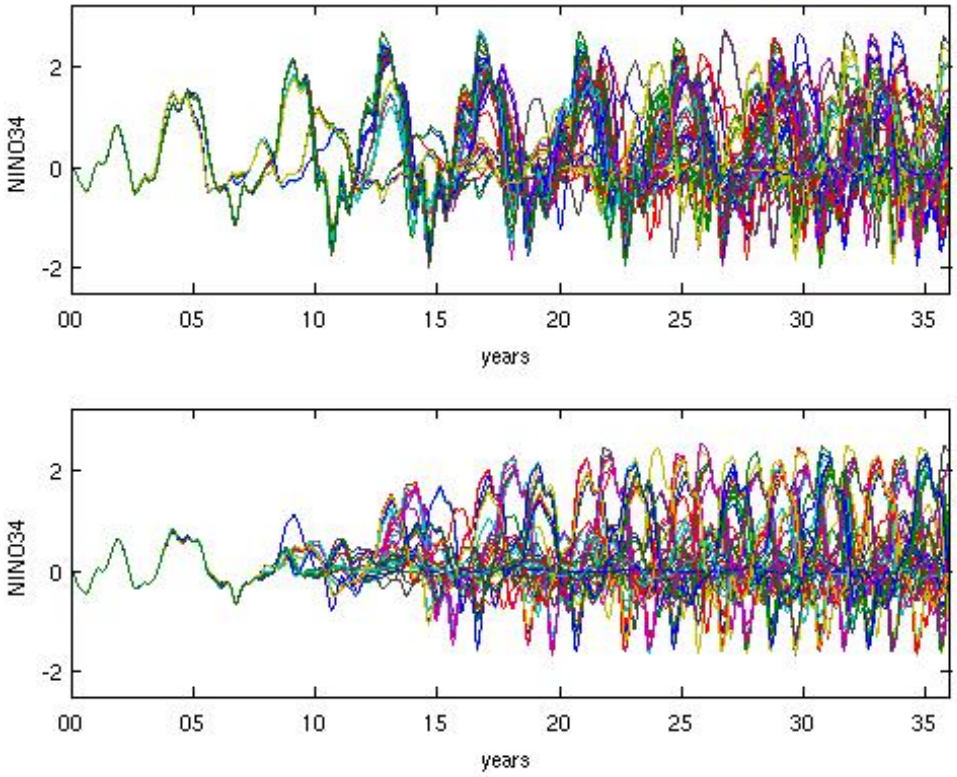


Figure C.1: Example of an ensemble forecast as a function of time for the 10-day model (top chart) and the 5-day model (bottom chart), experiment 6.E. Ensembles spread around different values of NINO34 index at different paces - quicker for the 10-day model and at different ranges - wider for the 10-day model.

C.2 Additional analysis for section 6.4.

In the IMS the Empirical Ignorance was occasionally found to be positive. For example, in years 2, 4-7, 13, 14, 18 or 19 the ensemble is far from the outcome (see Fig.C.2) which makes the Ignorance positive.

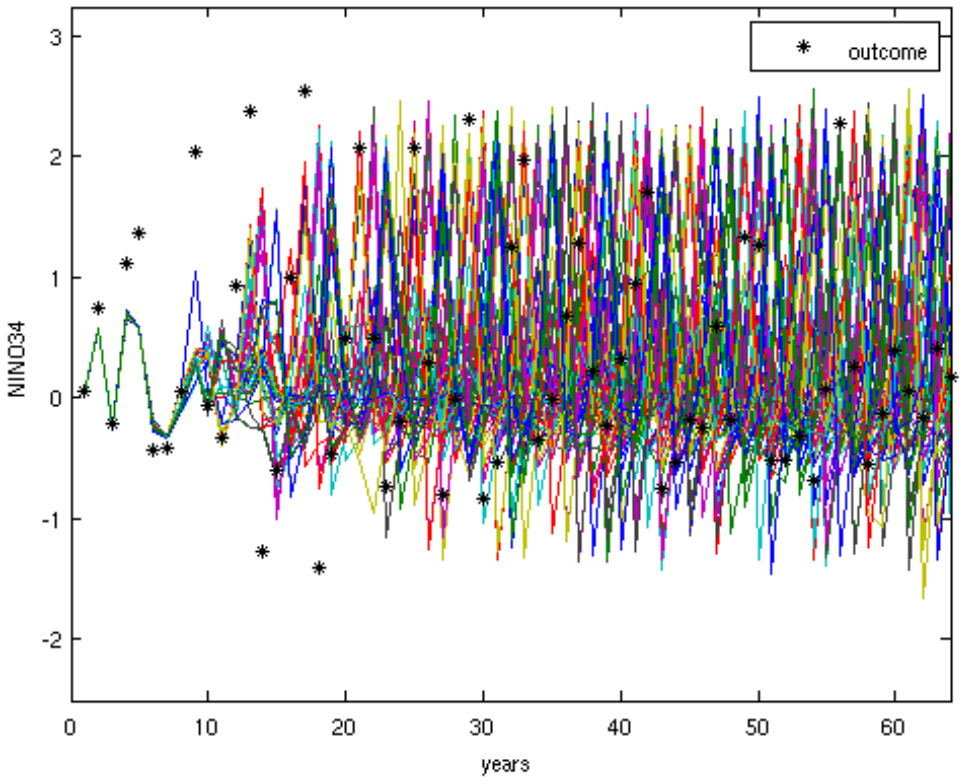


Figure C.2: An ensemble forecast shown as a function of time in the Imperfect Model scenario over 64 years in May along with outcome for each year. Each line shows a member of the ensemble while a black star illustrates the outcome in each year. In some years, the ensemble is far from the outcome, which results in the Empirical Ignorance being positive.

Bibliography

[1] Cane and Zebiak: Developing a Coupled ENSO Model. https://iridl.ldeo.columbia.edu/maproom/ENSO/New/canezebiak.html. Accessed: 25/04/2017.

[2] CMIP Coupled Model Intercomparison Project, CMIP3 - Overview. http://cmip-pcmdi.llnl.gov/cmip3_overview.html. Accessed: 19/03/2015.

[3] Ensemble forecasting. http://www.metoffice.gov.uk/research/weather/ensemble-forecasting. Accessed: 08/06/2017.

[4] Golden Gate Weather Services: El Niño and La Niña Years and Intensities. http://ggweather.com/enso/oni.htm. Accessed: 03/05/2017.

[5] Met Office, User guidance, What is climatology? https://www.metoffice.gov.uk/services/public-sector/contingency-planners/user-guidance. Accessed: 15/12/2017.

[6] NARCCAP North American Regional Climate Change Assessment Program. http://www.narccap.ucar.edu. Accessed: 19/03/2015.

[7] NOAA National Oceanic and Atmospheric Administration, U.S Department of Commerce. Climate data and reports. `http://www.noaa.gov/climate-data-and-reports`. Accessed: 19/03/2015.

[8] Reanalyses.org Home Page. `http://reanalyses.org`. Accessed: 03/01/2017.

[9] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium in Information Theory*, 19(6):276–281, Dec 1973.

[10] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.

[11] K. Anagnostopoulos. *Computational Physics, Vol II: A Practical Introduction to Computational Physics and Scientific Computing, Volume 2*. National Technical University of Athens, 2014.

[12] A. G. Bamston, M. Chelliah, and S. B. Goldenberg. Documentation of a highly ensorelated sst region in the equatorial pacific: Research note. *Atmosphere-Ocean*, 35(3):367–383, 1997.

[13] J. M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7:686–690, 1979.

[14] R. Binter. *Applied Probabilistic Forecasting*. PhD thesis, London School of Economics, 2011.

[15] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[16] G. Boeing. Visual analysis of nonlinear dynamical systems: Chaos, fractals, self-similarity and the limits of prediction. *Systems*, 4(4), 2016.

[17] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 2010.

[18] G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis Forecasting and Control*. Wiley, New Jersey, 2008.

[19] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678, 2008.

[20] A. D. Brunner. El Niño and world primary commodity prices: warm water or hot air? *Review of Economics and statistics*, 84(1):176–183, 2002.

[21] M. A. Cane, S. E. Zebiak, and S. C. Dolan. Experimental forecasts of El Niño. *Nature*, 321:827–832, 1986.

[22] S. Carl-Axel and S. von Holstein. Measurement of subjective probability. *Acta Psychologica*, 34(0):146 – 159, 1970.

[23] S. A. Changnon. Impacts of 1997-98 El Niño generated weather in the United States. *Bulletin of the American Meteorological Society*, 80(9):1819, 1999.

[24] C. Chatfield. *The Analysis of Time Series*. London: Chapman and Hall/CRC Press Company, 2004.

[25] M. Collins, R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A. Weaver, and M. Wehner. 12. long-term climate change: Projections, commitments and irreversibility. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on*

Climate Change, pages 1029–1136. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[26] H. Cramer. *Mathematical methods of statistics*. Princeton University Press, 1946.

[27] D. Demeritt, H. Cloke, F. Pappenberger, J. Thielen, J. Bartholmes, and M. Ramos. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*, 7(2):115–127, Jan. 2007.

[28] J. C. W. Denholm-Price. Can an ensemble give anything more than gaussian probabilities? *Nonlinear Processes in Geophysics*, 10(6):469–475, 2003.

[29] L. Descamps, C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cbron. Pearp, the mto-france short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, pages n/a–n/a, 2014.

[30] F. J. Doblas-Reyes, R. Hagedorn, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting ii. calibration and combination. *Tellus A*, 57(3):234–252, 2005.

[31] H. Du. *Combining statistical methods with dynamical insight to improve nonlinear estimation*. PhD thesis, London School of Economics, 2009.

[32] H. Du and L. Smith. Pseudo-Orbit Data Assimilation. Part I: The Perfect Model Scenario. *Journal of the Atmospheric Sciences*, 71:469–482, 2014.

[33] H. Du and L. Smith. Pseudo-Orbit Data Assimilation. Part II: Assimilation with Imperfect Models. *Journal of the Atmospheric Sciences*, 71:483–495, 2014.

[34] H. Du and L. A. Smith. Parameter estimation through ignorance. *Physical Review E*, 86(1):016213, 2012.

[35] E. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969.

[36] C. Ferro, D. Richardson, and A. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. *RMetS*, 15:1924, 1998.

[37] D. Friedman. Effective scoring rules for probabilistic forecasts. *Management Science*, 29(4), 1983.

[38] R. Frigg, S. Bradley, H. Du, and L. A. Smith. Laplace's Demon and the Adventures of His Apprentices, January 2014.

[39] R. Frigg and J. Reiss. The philosophy of simulation: hot new issues or same old stew? *Synthese*, 169:593613, 2009.

[40] I. Gilmour. *Nonlinear Model Evaluation: Iota -shadowing, Probabilistic Prediction and Weather Forecasting.* University of Oxford, 1998.

[41] M. Glantz. Floods, fire, and famine; Is El Niño to Blame? *Oceanus*, 27(2):14–19, 1984.

[42] M. H. Glantz. *Currents of Change: Impacts of El Niño and La Niña on Climate and Society.* Cambridge University Press, Cambridge, UK, 2001.

[43] M. H. Glantz, R. W. Katz, and N. Nicholls. *Teleconnections Linking Worldwide Climate Anomalies.* Cambridge University Press, 2009.

[44] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[45] T. Gneiting, A. H. Westveld, A. E. Raftery, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Technical report, Monthly Weather Review, 2005.

[46] L. Goddard and M. Dilley. El Niño: Catastrophe or Opportunity. *Journal of Climate*, 18(5):651–665, 2005.

[47] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14:107–114, 1952.

[48] M. S. Halpert and C. F. Ropelewski. Surface Temperature Patterns Associated with the Southern Oscillation. *Journal of Climate*, 5(6):577–593, 1992.

[49] M. Henon. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50(1):69–77, 1976.

[50] M. Hollander, D. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. John Wiley Sons, New York, USA, 2014.

[51] S. M. Hsiang, K. C. Meng, and M. A. Cane. Civil conflicts are associated with the global climate. *Nature*, 476(7361):438–441, 2011.

[52] T. Iizumi, J. Luo, A. J. Challinor, G. Sakurai, M. Yokozawa, H. Sakuma, M. E. Brown, and T. Yamagata. Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nature Communications*, 5(3712), 2014.

[53] K. Ikeda. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, opt. *Comm*, pages 257–261, 1979.

[54] G. James, T. Hastie, D. Witten, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, New York, USA, 2015.

[55] A. Jarman. *On the provision, reliability, and use of hurricane forecasts on various timescales.* PhD thesis, London School of Economics, 2014.

[56] J.Bröcker and L. Smith. Scoring probabilistic forecasts: the importance of being proper. *Tellus A*, 22(2), 2007.

[57] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 90, 1995.

[58] K. Judd and T. Stemler. Forecasting: It is not about statistics, it is about dynamics. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 368(1910):263–71, 2010.

[59] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.

[60] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[61] G. N. Kiladis and H. F. Diaz. Global Climatic Anomalies Associated with Extremes in the Southern Oscillation. *Journal of Climate*, 2(9):1069–1090, 1989.

[62] R. S. Kovats, M. J. Bouma, S. Hajat, E. Worrall, and A. Haines. El Niño and health. *The Lancet*, 362(9394):1481–1489, 2003.

[63] C. E. Leith. Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102:409, 1974.

[64] M. Leutbecher and T. N. Palmer. Ensemble forecasting. *J. Comput. Phys.*, 227(7):3515–3539, Mar. 2008.

[65] E. N. Lorenz. Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, 20(2):130–141, Mar. 1963.

[66] R. Machete and L. Smith. Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A*, 68(29393), 2016.

[67] C. MacKay and J. David. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[68] J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

[69] L. O. Mearns, R. Arritt, S. Biner, M. S. Bukovsky, S. McGinnis, S. Sain, D. Caya, J. C. Jr., D. Flory, W. Gutowski, E. S. Takle, R. Jones, R. Leung, W. Moufouma-Okia, L. McDaniel, A. M. B. Nunes, Y. Qian, J. Roads, L. Sloan, and M. Snyder. The north american regional climate change assessment program: Overview of phase i results. *Bulletin of the American Meteorological Society*, 93(9):1337–1362, 2012.

[70] L. O. Mearns, W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes, and Y. Qian. A Regional Climate Change Assessment Program for North America. *Eos, Transactions American Geophysical Union*, 90(36):311–311, 2009.

[71] F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119, 1996.

[72] H. Murphy. On the ranked probability score. *Journal of Applied Meteorology*, 8(6):988–989, 1969.

[73] J. Murphy. The impact of ensemble forecasts on predictability. *Quarterly Journal of the Royal Meteorological Society*, 114(480):463–493, 1988.

[74] K. Murphy. *Machine Learning: A Probabilistic Perspective*. Massachusetts Institute of Technology, 2012.

[75] N. Nakicenovic, J. Alcamo, A. Grubler, K. Riahi, R. Roehrl, H.-H. Rogner, and N. Victor. *Special Report on Emissions Scenarios (SRES), A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2000.

[76] J. E. Oliver. *The Encyclopedia of World Climatology (Encyclopedia of Earth Sciences Series)*. Springer, Aug. 2005.

[77] E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, Cambridge, 2002.

[78] T. Palmer. Predictability of weather and climate: from theory to practice–from days to decades. In *Realizing Teracomputing: Proceedings of the Tenth ECMWF Workshop on the Use of High Performance Computing in Meteorology: Reading, UK, 4-8 November, 2002*. World Scientific, 2003.

[79] W. S. Parker. Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Modern Physics*, 41:263272, 2010.

[80] E. Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[81] H.-O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals: New Frontiers of Science*. Springer, Feb. 1993.

[82] P. Peng, A. Kumar, H. van den Dool, and A. G. Barnston. An analysis of multimodel ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research: Atmospheres*, 107(D23):ACL 18–1–ACL 18–12, 2002.

[83] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes. The Art of Scientific Computing.* Cambridge University Press, 2007.

[84] D. S. Richardson. Predictability and economic value. In T. Palmer and R. Hagedorn, editors, *Predictability of Weather and Climate*, pages 628–644. Cambridge University Press, 2006. Cambridge Books Online.

[85] R. Rohli and A. Vega. *Climatology.* Cambridge University Press, 2011.

[86] C. F. Ropelewski and M. S. Halpert. Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, 115(8):1606–1626, 1987.

[87] C. F. Ropelewski and M. S. Halpert. Precipitation Patterns Associated with the High Index Phase of the Southern Oscillation. *Journal of Climate*, 2(3):268–284, 1989.

[88] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002.

[89] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55(1):16–30, 2003.

[90] E. S. Sarachik and M. A. Cane. *The El Niño-Southern Oscillation Phenomenon*. Cambridge University Press, Cambridge, UK, 2010.

[91] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Apr. 1986.

[92] S. Singh, M. R. James, and M. R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 512–519, Arlington, Virginia, United States, 2004. AUAI Press.

[93] L. Smith. What might we learn from climate forecasts? *Proc. National Acad. Sci. USA*, 99(4):2487–2492, 2002.

[94] L. Smith and J. Hansen. Extending the limits of forecast verification with the minimum spanning tree. *Monthly Weather Review*, 132(6):1522–1528, 2004.

[95] L. Smith and K. Judd. Indistinguishable states i. perfect model scenario. *Physica D*, 151:125–141, 2001.

[96] L. Smith and K. Judd. Indistinguishable states ii. imperfect model scenario. *Physica D*, 196:224–242, 2004.

[97] L. A. Smith. Accountability and error in ensemble forecasting. *1995 ECMWF Seminar on Predictability,Reading,UK*, 1:351–368, 1995.

[98] L. A. Smith. Disentangling uncertainty and error: On the predictability of nonlinear systems. In A. I. Mees, editor, *Nonlinear Dynamics and Statistics*, pages 31–64. Birkhäuser Boston, 2000.

[99] L. A. Smith. Predictability past, predictability present. In T. Palmer and R. Hagedorn, editors, *Predictability of Weather and Climate*, pages 217–250. Cambridge University Press, 2006. Cambridge Books Online.

[100] L. A. Smith. *Chaos: A Very Short Introduction*. Oxford University Press, 2007.

[101] L. A. Smith, E. A. Sienkiewicz, and E. Wheatcroft. Distinguishing inadequacy from sensitivity in simulation based forecasting: a guide to resource allocation. *In preperation*, 2017.

[102] L. A. Smith, E. B. Suckling, E. L. Thompson, T. Maynard, and H. Du. Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, 132:31 45, 2015.

[103] L. A. Smith, C. Ziehmann, and K. Fraedrich. Uncertainty dynamics and predictability in chaotic systems. *Quarterly Journal of the Royal Meteorological Society*, 1:351–368, 1999.

[104] P. Smith. *Explaining Chaos*. Cambridge University Press, Cambridge, 1998.

[105] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, 2003.

[106] T. Stocker, D. Qin, G.-K. Plattner, L. Alexander, S. Allen, N. Bindoff, F.-M. Breon, J. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J. Gregory, D. Hartmann, E. Jansen, B. Kirtman, R. Knutti,

K. Krishna Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G. Meehl, I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L. Talley, D. Vaughan, and S.-P. Xie. Technical summary. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, book section TS, pages 33–115. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[107] S. H. Strogatz. *Nonlinear Dynamics and Chaos.* Oxford University Press, Oxford, 2000.

[108] E. Thompson, E. A. Sienkiewicz, and L. A. Smith. When (not) to downscale. *In preperation*, 2017.

[109] Z. Toth and E. Kalnay. Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74:2317–2330, Dec. 1993.

[110] A. Tsonis. *Chaos: from theory to applications.* Plenum Press, New York, 1992.

[111] D. Wilks. *Statistical Methods in the Atmospheric Sciences.* Academic Press, London,UK, 2006.

[112] R. Winkler and A. Murphy. Good probability assessors. *Journal of applied meteorology*, 7:751 758, 1968.

[113] X.Wang and C.H.Bishop. Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteorol. Soc.*, 10(131):965–986, 2005.

[114] S. E. Zebiak and M. A. Cane. A model El Niño-Southern Oscillation. *Monthly Weather Review*, 115:2262–2278, 1987.